·热点与综述· 文章编号：1000-3428(2021)09-0018-16 文献标志码：A 中图分类号：TP391

知识感知的预训练语言模型综述

李瑜泽[1,2,3]，栾 馨[1,2,3]，柯尊旺[2,3,4]，李 哲[2,3,4]，吾守尔·斯拉木[1,2,3]
（1.新疆大学 信息科学与工程学院，乌鲁木齐 830046；2.新疆多语种信息技术实验室，乌鲁木齐 830046；
3.新疆多语种信息技术研究中心，乌鲁木齐 830046；4.新疆大学 软件学院，乌鲁木齐 830046）

摘 要：随着自然语言处理(NLP)领域中预训练技术的快速发展，将外部知识引入到预训练语言模型的知识驱动方法在NLP任务中表现优异，知识表示学习和预训练技术为知识融合的预训练方法提供了理论依据。概述目前经典预训练方法的相关研究成果，分析在新兴预训练技术支持下具有代表性的知识感知的预训练语言模型，分别介绍引入不同外部知识的预训练语言模型，并结合相关实验数据评估知识感知的预训练语言模型在NLP各个下游任务中的性能表现。在此基础上，分析当前预训练语言模型发展过程中所面临的问题和挑战，并对领域发展前景进行展望。

关键词：自然语言处理；知识表征；语义知识；预训练；语言模型

开放科学(资源服务)标志码(OSID)：

中文引用格式：李瑜泽，栾馨，柯尊旺，等.知识感知的预训练语言模型综述[J].计算机工程，2021，47(9)：18-33.
英文引用格式：LI Y Z，LUAN X，KE Z W，et al.Survey of knowledge-aware pre-trained language models[J].Computer Engineering，2021，47(9)：18-33.

Survey of Knowledge-Aware Pre-Trained Language Models

LI Yuze[1,2,3]，LUAN Xin[1,2,3]，KE Zunwang[2,3,4]，LI Zhe[2,3,4]，Wushour Silamu[1,2,3]
（1.College of Information Science and Engineering，Xinjiang University，Urumqi 830046，China；2.Xinjiang Laboratory of Multi-Language Information Technology，Urumqi 830046，China；3.Xinjiang Multi-Language Information Technology Research Center，Urumqi 830046，China；4.School of Software，Xinjiang University，Urumqi 830046，China）

【Abstract】In the field of Natural Language Processing(NLP)，the recent years has witnessed a rapid development in pre-training technology，and the knowledge-driven method that injects external knowledge into a pre-trained language model performs excellently in NLP tasks.The techniques of knowledge representation learning and pre-training provide theoretical foundation for the knowledge-based pre-training method.This paper briefly introduces the development of the classical pre-trained methods. Then it analyzes the representative knowledge-aware pre-trained language models supported by new pre-training technology.According to the types of external knowledge，this paper introduces the pre-trained language models injected with different external knowledge.Based on relevant experimental data，it subsequently evaluates the performance of the knowledge-aware pre-trained language models in various downstream tasks of NLP.On this basis，the paper analyzes the problems and challenges faced by the developing pre-trained language models，and discusses the development trends of this field.

【Key words】Natural Language Processing(NLP)；knowledge representation；semantic knowledge；pre-training；language model
DOI：10. 19678/j. issn. 1000-3428. 0060823

0 概述

自然语言处理(Natural Language Processing，NLP)是人工智能和语言学领域的分支学科，近年来，NLP的发展取得了巨大进步，任务划分也更加详尽，其主要下游任务，如关系抽取[1]、文本分类[2]、对话生成、机器翻译和共指消解[3]等均离不开预训练技术。预训练技术使用大规模无标注的文本语料库

基金项目：国家重点研发计划(2017YFC0820702-3)。
作者简介：李瑜泽(1997—)，女，硕士研究生，主研方向为自然语言处理、舆情分析；栾 馨，硕士研究生；柯尊旺(通信作者)，讲师、博士；李 哲，硕士研究生；吾守尔·斯拉木，教授、博士生导师。
收稿日期：2021-02-05 **修回日期：**2021-03-31 **E-mail：**kzwang@xju.edu.cn

对深层网络结构进行训练，从而得到一组模型参数，这种深层网络结构通常被称为“预训练模型”[4]。将训练得到的模型参数运用于后续特定的下游任务，可避免从零开始进行再训练的繁琐过程。预训练模型的优点在于训练代价较小，配合下游任务能够达到更好的收敛效果，可以较好地提升模型的性能。

自BERT[5]模型被提出以来，各种预训练语言模型层出不穷。然而，非知识感知的预训练语言模型通常采用浅层网络结构，只进行简单的学习，无法理解更高层次的文本概念，如语义角色、指代等；而知识感知的预训练语言模型能够根据不同的应用目的而加强不同的模块，使模型具备更加专业的业务能力。因此，面对智能要求和专业化程度更高的深层次NLP任务，需要探索并分析将特定种类的外部知识嵌入到特定用途的预训练语言模型中。

目前，已有部分研究人员尝试将外部知识嵌入到预训练语言模型中，评估其在特定下游任务中的性能表现。NLP的外部知识含义广泛，涵盖了语言学知识、语义知识、常识知识、事实知识以及领域知识等[6]。知识丰富的预训练语言模型通常从通用的大规模文本语料库中学习通用的语言表示，以灵活应对多种下游任务并进行相应的微调[7]。基于知识感知[8]对预训练语言模型进行扩展，使得预训练技术进入了一个新的发展阶段。

本文在分析预训练技术发展历程中经典语言模型的基础上，将研究重点转向知识感知的预训练语言模型，分析此类模型相比其他预训练语言模型的优势，研究知识感知对于预训练语言模型性能提升的增强作用，并在多种知识驱动的NLP下游任务中评估模型的性能表现。同时，搜集最新知识感知的预训练语言模型的相关研究成果，对其中多种模型对比实验的结果加以分析，从中归纳总结已有知识感知的预训练语言模型适用于何种NLP下游任务，从而对未来的研究方向进行展望。

1　研究背景及相关技术

1.1　知识表示学习技术

知识感知的预训练语言模型大多基于知识图谱引入丰富的外部知识，因此，本节介绍知识表示学习的原理和其中具有代表性的相关方法。

知识图谱是一种由实体节点和关系组成的语义网络。表示学习旨在将研究对象的语义信息表示为稠密低维实值向量，而知识表示学习则在词向量的启发下面向知识图谱中实体和关系的语义信息进行学习，通过将实体和关系映射到连续的低维向量空间中以实现知识的有效表示[9]。

知识的表示方法主要有符号表示[10]和数值表示：符号表示通常包含字符、关联图和符号等；数值表示通常用标量、向量、概率分布等数值表达事实与知识。

TransE模型[11]是传统且具有代表性的知识表示学习方法，随后针对TransE模型进行改进的方法也层出不穷。TransE模型主要考虑向低维向量空间中嵌入实体和多关系数据的关系问题，采用基于翻译[12]的思想，通过将知识图谱中的关系解释为对应于实体的某种低维嵌入平移向量[13]，从而实现对关系的建模，因此，TransE也被称作翻译模型。将三元组(h,r,t)看作头实体h到尾实体t利用关系r所进行的翻译，则TransE的损失函数如下：

$$f_r(h,t)=\|h+r-t\|_{l_1/l_2} \tag{1}$$

其中：l_1和l_2表示向量空间。在TransE的改进模型中，性能较好的PTransE模型通过对关系路径进行编码，实现了实体和关系在低维向量空间中的嵌入，其能够利用关系路径约束资源配置并衡量关系路径的可靠性。

此外，还有诸多传统的知识表示学习方法及其改进模型。例如，将不同的实体和关系对象映射到独立语义空间的TransR模型、注重实体与关系之间交互的TransD模型、专注解决多重关系语义问题的TransG模型以及用来处理复杂关系映射问题的TransH模型等。

1.2　预训练技术

预训练技术利用大型语料库学习通用语义表示，从而将预训练好的模型的相应结构和权重直接应用于下游任务，其借鉴了迁移学习[14]的思想。迁移学习的本质是在一个数据集上训练基础模型，通过微调等方式使模型可以在其他不同的数据集上处理不同的任务。

1.3　知识融合的预训练方法

1.3.1　语言建模的初步知识

本文在不失一般性的前提下研究自回归语言建模问题，将文本X视为符号序列（单词或子单词）：$X=\{\omega_1,\cdots,\omega_i,\cdots,\omega_n\}$，语言概率的经典单向因式分解如下：

$$p(X)=\prod_i p(\omega_i\mid\omega_{<i}) \tag{2}$$

其中：$\omega_{<i}$指i之前出现的所有标记。上述条件概率可以采取各种方式进行参数化，一种有效的方法是使用单向变压器（如GPT-2）来实现：

$$p(\omega_i\mid\omega_{<i})=\text{transformer}(\omega_i\mid\omega_{<i}) \tag{3}$$

由式（3）可知，可以通过增大Transformer参数的方法促使语言建模能力提升，即随着Transformer层的扩大和加深，Transformer语言模型逐渐能够输出词法和语法模式之外的更加复杂的语义。

1.3.2　知识与预训练的融合方式

预训练语言模型中的所有语义都是由Transformer间接捕获的，预训练中也并无明确的要求用以更好

地捕捉知识，而当前期望是其能含蓄地捕捉原始单词序列下的知识，方法之一是采用知识感知的语言建模（KALM）框架[15]，在相同数量的Transformer参数中封装更多的信息，而非叠加更多的层或添加外部知识存储。

知识与预训练融合的第一步是利用实体标记器，通过形成一个额外的实体标记序列表明预训练过程的输入和输出中存在实体。在知识感知输入阶段，首先每个单词标记都有一个输入嵌入，同时允许模型学习每个实体的一个实体嵌入，然后将2种嵌入相结合形成知识感知输入，这里的所有嵌入都是随机初始化并在训练前进行学习的。在知识感知输出阶段，使用一个输出头来表示单词概率，另一个输出头来表示实体，并共享单词和实体之间的所有Transformer层。最后，将知识感知的输入和输出引入到标准的多任务设置中，结合语言建模的损失并引入超参数，平衡标准预训练语言模型中的交叉熵与实体预测损失之间的关系。通过上述过程，知识与预训练的融合已基本实现。

2 预训练技术的发展历程

预训练的思想起初来源于计算机视觉领域，鉴于预训练在计算机视觉领域取得的较好效果，NLP开始尝试使用预训练技术实现迁移学习。当预训练技术运用于NLP领域时，训练得当的语言模型可以捕获与下游任务相关的许多知识，如长期依赖、层次关系等。此外，在NLP领域进行训预练的显著优势是训练数据可以来源于任意无监督文本语料，即训练过程拥有无限量的训练数据。

早期的预训练是一种静态技术，NNLM[16]作为使用神经网络实现语言模型的经典范例，早在2003年就已被提出，Word2Vec[17]借鉴NNLM的思想使用语言模型得到词向量，随后GloVe[18]和FastText[19]相继被提出，使得这种静态的预训练技术逐渐成为最常用的文本表征技术之一[20]。但是，这种预训练技术归根到底是一种静态技术，其对不同语境的适应能力较差，且对下游任务的性能提升收效甚微。

为解决上述问题，研究人员开始关注动态的预训练技术。自2018年PETERS等[21]提出一种上下文相关的文本表示方法以来，DPT、BERT、XLNet[22]等预训练语言模型相继被提出，预训练技术逐渐在NLP领域得到广泛应用。BERT的出现将预训练语言模型的发展推向高潮，此后相继涌现出一系列新式的预训练技术，就模型结构而言，它们所训练的模型一类是基于BERT的改进模型，另一类则是XLNet，关于预训练技术的研究自此进入新阶段。

2.1 静态预训练技术

本节主要总结分析NNLM、Word2Vec、GloVe、FastText等静态预训练技术。

2003年BENGIO提出的NNLM是早期使用神经网络实现语言模型的经典模型，其本质是一个N-Gram语言模型。研究人员将模型的第一层特征映射矩阵当作词的文本表征，从而开创将词语表示为向量形式的模式，启发了后续Word2Vec的有关研究。NNLM的不足之处在于只利用了上文信息，并未结合更多的上下文信息，此外，其输出层存在词表较大引发计算量过大的问题。

2013年MIKOLOV提出的Word2Vev在模型结构和训练技巧方面对NNLM进行了优化，并提出使用语言模型作为词向量的构思。该文首先提出CBOW[23]和Skip-gram[24]以简化模型结构，其次使用Hierarchical Softmax[25]和Negative Sampling技术解决普通Softmax计算量过大的问题，最后将词向量更换为需要由上下文词向量[26]求和得到的语境向量。上述优化使得在大规模无监督文本语料上训练得到的词向量在语义上取得了良好的性能表现。

2014年PENNINGTON等提出的GloVe是一个基于全局词频统计的词表征工具，它可以将某个单词表达成由实数组成的向量，通过向量获取单词之间的语义特性，对向量进行运算可以得出2个单词之间的语义相似度[27]。与Word2Vec相比，GloVe主要利用词语的共现信息构建模型，其训练主要包含统计共现矩阵和训练获取词向量2个步骤。

2016年，Word2Vec的提出者MIKOLOV再次提出FastText，其利用有监督标记的文本分类数据进行训练。FastText的网络结构与CBOW类似，但学习目标是人工标注的分类结果，两者均使用Hierarchical Softmax技术加速训练。此外，FastText为了利用更多的语序信息，增加了N-Gram[28]的输入信息，在训练词向量时引入subword来处理长词，这些改进均使得FastText的训练速度大幅提升。

2.2 动态预训练技术

随着NLP应用场景的不断丰富，以往采用静态预训练技术训练的模型有时无法很好地完成特定的任务。在此情形下，动态预训练技术方案应运而生，本节主要概述其中经典且具有代表性的ELMo、GPT和BERT。

2018年，ELMo提出一种上下文相关的文本表示方法，根据当前上下文对Word Embedding[29]进行动态调整，其任务采用典型的两阶段划分：第一阶段利用语言模型进行预训练；第二阶段在下游任务进行时从预训练网络中提取对应单词，将其网络各层的Word Embedding作为新特征补充到下游任务中。虽然ELMo在当时的适用范围非常广泛，但其选择LSTM[30]作为特征提取器，而非提取能力更强的Transformer，导致在某些任务中的表现受到一定限制。

2018年，RADFORD等提出的GPT使用生成式

方法训练语言模型,选用Transformer作为特征提取器,同样也划分为两阶段过程:第一阶段利用语言模型进行预训练,但其采用的是单向语言模型;第二阶段通过微调(Fine-tuning)模式处理下游任务。由于仅选取预测词的上文进行预测,因此GPT在某些应用任务中仍然无法取得较好的效果。

2018年,谷歌团队的BERT模型一经发布,立即刷新了11项NLP任务的性能记录。BERT模型的架构基于Transformer,实现多层双向的Transformer编码器,其同样被划分为两阶段过程:第一阶段的预训练过程包含Masked Language Model[31]和Next Sentence Prediction[32]2个任务;第二阶段使用Fine-tuning模式处理下游任务。与GPT相比,BERT效果更优,这主要归功于双向语言模型发挥的主要作用,尤其在下游任务中,其作用更加凸显。此外,Next Sentence Prediction在个别任务中也有出色表现。虽然BERT在众多任务中表现优异,但其参数量巨大,导致计算资源消耗大幅提升。

2.3 新兴预训练技术

新兴预训练技术从模型结构上可分为两类:一类是能够获得真双向上下文信息的自回归语言模型XLNet;另一类是基于BERT的改进模型,包括改进生成任务、引入多任务、引入知识以及改进训练方法等改进模型。

在BERT的众多改进模型中,MASS[33]和UNILM[34]2个模型均致力于改进BERT在其生成任务中的表现。其中,MASS在机器翻译任务中效果显著,UNILM则在摘要生成、问题生成、对话回复生成以及生成式问答任务中效果更佳。MT-DNN[35]和ERNIE 2.0[36]2个模型通过引入多任务学习改进BERT。其中,MT-DNN能更好地在小数据集上实现微调,ERNIE 2.0则在阅读理解、情感分析以及问答任务上效果显著。ERNIE 1.0和ERNIE(THU)[37]2个模型通过引入知识,使预训练语言模型能够学习到蕴含在深层文本中的潜在知识。其中,ERNIE 1.0在自然语言推断[38]、语义相似度、情感分析、命名实体识别[39]以及问答匹配任务中较BERT表现更佳,ERNIE(THU)则在知识驱动型任务中效果显著。RoBERTa[40]模型采用比BERT更加精细的参数和训练方法,在句子关系推断任务中具有比BERT更好的性能。BERT-WWM更改原预训练阶段的训练样本生成策略,当被掩蔽的词块标记属于一个整词时,所有构成一个完整词的词块标记将一起被掩蔽。由于BERT-WWM缺乏中文语言相关模型,哈工大讯飞实验室提出了全词覆盖的中文BERT预训练模型,即BERT-wwm-ext[41],其增加了训练数据集同时也增加了训练步数,进一步促进中文信息处理研究的发展。为解决目前预训练模型参数量过大的问题,ALBERT(A Lite BERT)[42]提出2种能够大幅减少预训练模型参数量的方法,用SOP(Sentence-Order Prediction)任务代替BERT中的NSP(Next-Sentence Prediction)任务,在多个自然语言理解任务中取得了较优的结果。在BERT模型的基础上进行改进得到的分词级别的预训练模型SpanBERT[43],对随机的邻接分词而非随机的单个词语添加掩模,通过使用分词边界的表示来预测被添加掩模的分词内容,其在问答、指代消解等分词选择任务中取得了较大的进展。华为、华科联合提出的TinyBERT模型[44]是一种基于transformer而专门设计的知识蒸馏[45]方法,其模型大小不足BERT的1/7,但将处理速度相对BERT提高了9倍,大幅降低了其应用成本。

XLNet模型使用排列语言模型、双流自注意力和循环三大机制,不仅解决了BERT单词之间预测不独立的问题,还使得自回归模型也可获得真双向的上下文信息。

预训练的语言表示模型(如ELMo、BERT和XLNet等)从大规模的非结构化和无标记语料库中学习有效的语言表示,并凭借其优异的性能在各种NLP任务中发挥着重要作用。然而,它们大多缺乏真实的世界知识,最新的研究结果表明,可以通过将外部知识注入到预训练语言模型中来解决某些文本理解较为局限的问题。

3 知识感知的预训练语言模型

预训练语言模型通常从通用的大规模文本语料库中学习通用的语言表示,但其大多缺乏特定领域的知识。目前已有许多研究人员尝试将外部知识库中的领域知识整合到预训练语言模型中,再将这些模型分别运用于各自适用的NLP下游任务并从中监测其表现,结果表明,这种通过知识嵌入增强预训练语言模型性能的方法具有有效性。

本文根据外部知识的范围,从语言学、语义、常识、事实以及特定领域的知识等方面出发,对现有相关模型进行归纳和总结,重点介绍其中具有代表性的知识感知的预训练语言模型。

3.1 事实知识融合的预训练语言模型

现有的预训练语言模型很少考虑引入知识图谱(KGs)[46],而知识图谱可以为语言理解提供丰富的结构化知识事实。

清华大学与华为公司的相关研究人员认为,KGs中的信息实体可以作为外部知识来增强语言表示,因此,他们提出一种使用信息实体增强语言表示的模型ERNIE(THU),该模型利用大规模的文本语料库和KGs训练得到一个增强的语言表示模型。

3.1.1 ERNIE(THU)总体架构

如图1所示,ERNIE(THU)模型架构由2个堆叠的模块组成:

1)底层文本编码器(T-Encoder)负责从输入标记

中捕获基本的词汇和语法信息。

2)上层知识丰富的编码器(K-Encoder)负责将额外的面向标记的知识信息集成到底层的文本信息中,以便在统一的特征空间中表示标记和实体的异构信息。

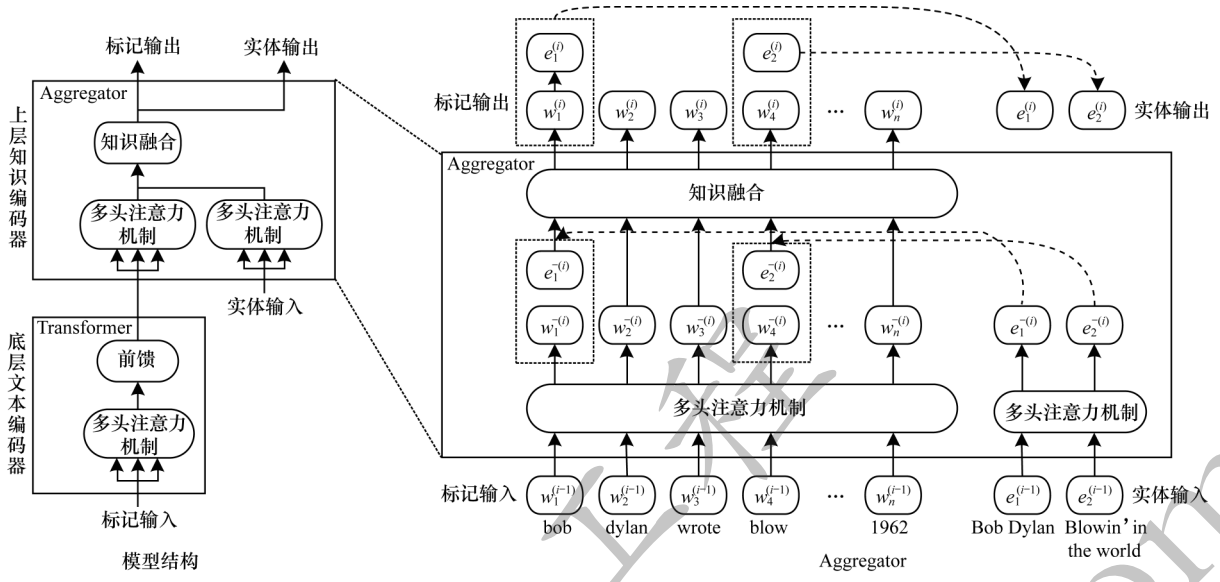

图1 ERNIE(THU)总体架构

Fig.1 The architecture of ERNIE(THU)

3.1.2 ERNIE(THU)预训练

与BERT相似,ERNIE(THU)也采用掩码语言模型(MLM)和下一句预测(NSP)作为预训练任务,以从文本标记中捕捉词法和句法信息。

为能够通过信息实体将知识注入到语言表示中,可以采用随机掩盖部分标记-实体对齐的方法,要求系统基于对齐的标记预测对应的实体,且只需根据给定的实体序列来预测实体,而非KGs中的所有实体。上述过程称为去噪实体自动编码器(dEA)。

给定标记序列$\{w_1,w_2,\cdots,w_n\}$及其对应的实体序列$\{e_1,e_2,\cdots,e_m\}$,其中,n为令牌序列的长度,m为实体序列的长度。为标记w_i定义对齐的实体分布如下:

$$p(e_j|w_i)=\frac{\exp(\text{linear}(w_i^o)\cdot e_j)}{\sum_{k=1}^{m}\exp(\text{linear}(w_i^o)\cdot e_k)} \tag{4}$$

其中:linear(·)为线性层。式(4)用来计算dEA的交叉熵损失函数,预训练损失是dEA、MLM和NSP的损失总和。

3.1.3 ERNIE(THU)微调

对于常见的NLP任务,ERNIE(THU)采用类似于BERT的微调过程,而对于知识驱动的NLP任务(如关系分类和实体键入),ERNIE(THU)设计了特殊的微调过程。

关系分类任务要求系统根据上下文对给定实体对的关系标签进行分类,通过添加2个标记符号修改输入标记序列以突出实体提及,这些额外的标记符号在传统关系分类模型中扮演着类似于位置嵌入的角色。ERNIE(THU)同样也采取[CLS]标记嵌入分类,还分别为头部实体和尾部实体设计了不同的标记[HD]和[TL]。

实体键入任务的特定微调是关系分类任务微调的简化过程,由于之前的类型模型充分利用了上下文嵌入和实体提及嵌入,因此修改后的输入序列带有提及标记符[ENT],可以引导ERNIE(THU)将上下文信息和实体提及信息相结合。

3.1.4 ERNIE(THU)评估

在FIGER和Open Entity 2个数据集上对BERT和ERNIE(THU)进行微调,评估两者在实体键入任务中的性能;在FewRel和TACRED 2个数据集上对BERT和ERNIE(THU)进行微调,评估两者在关系分类任务中的性能。实验结果如表1所示。

表1 BERT和ERNIE(THU)的性能对比结果

Table 1 Performance comparison results of BERT and ERNIE(THU) %

数据集	指标	BERT	ERNIE(THU)
FIGER	Acc	52.04	57.19
Open Entity	F1	76.37	78.42
FewRel	F1	84.89	88.32
TACRED	F1	66.00	67.97

从表1可以看出,ERNIE(THU)模型能同时充分利用词汇、句法和知识信息,在2个知识驱动的NLP任务,即实体键入和关系分类上取得优于BERT的性能。但是,ERNIE(THU)仍存在一定的局限性,比如受命名实体识别(NER)的影响,其依赖于NER提取的准确度,模型复杂度过高。

3.1.5　其他事实知识融合的预训练语言模型

ERNIE(THU)模型将预训练在KGs上的实体嵌入与文本中提到的相应实体相结合以增强文本表示。与ERNIE(THU)类似，KG-BERT[47]模型结合预训练模型BERT并修改其输入使BERT适配知识库三元组的形式，以便使更丰富的上下文表示与模型相结合，促使其在三元组分类、链接预测以及关系预测等任务中达到SOTA效果。KnowBERT[48]将BERT与实体链接模型相联合，以端到端的方式合并实体表示：使用常识知识库ConceptNet[49]和情感词典NEC_VAD作为外部知识来源，采用动态上下文感知情感图注意力机制，计算每个token融入知识后的上下文表示；引入相关性因子和情感因子2个因素来衡量上下文相关度和情感强度的权重；提出一种层次化自注意力机制，在利用对话结构表示形式的同时学习上下文语言的向量表示形式。KEPLER[50]将知识嵌入和语言建模目标相结合并进行优化：通过添加类似TransE的预训练机制以增强相应的文本表示，将每个实体与对应的维基百科描述相链接，使之均获得对应的文本描述信息；对于每一个三元组<头实体，关系，尾实体>，采用基于BERT的编码器，利用实体的描述信息对每个实体进行编码；在得到头实体和尾实体对应的表示之后，基于头实体和关系预测尾实体。上述模型都是通过实体嵌入的方式注入KGs的结构信息。

3.2　领域知识融合的预训练语言模型

现有预训练语言模型大多可以在大规模开放领域语料库上进行预训练以获得一般的语言表示，然后在特定的下游任务中进行微调以吸收特定领域的知识。但是，由于预训练和微调之间可能存在领域差异，此类模型往往在知识驱动任务中表现不佳。为此，文献[51]提出基于KGs知识支持的语言表示模型K-BERT，该模型将从KGs中提取的相关三元组作为领域知识显式地注入到句子中，得到BERT的扩展树形输入。由于K-BERT能够从预训练的BERT中加载模型参数，因此无须自行进行预训练，只需配备KGs即可将领域知识注入到模型中。

3.2.1　K-BERT总体架构

如图2所示，K-BERT模型架构由知识层、嵌入层、可见层、掩模转换器等4个模块组成。对于输入的句子，知识层首先从KGs中向其注入相关三元组，将原句转化为知识丰富的句子树；随后句子树被同时输送到嵌入层和可见层，转化为一个标记级别的嵌入表示和一个可见矩阵，可见矩阵用来控制每个标记的可见区域，防止因注入过多知识而导致原句句意改变；最后通过掩模转换器编码后用于下游任务的输出。

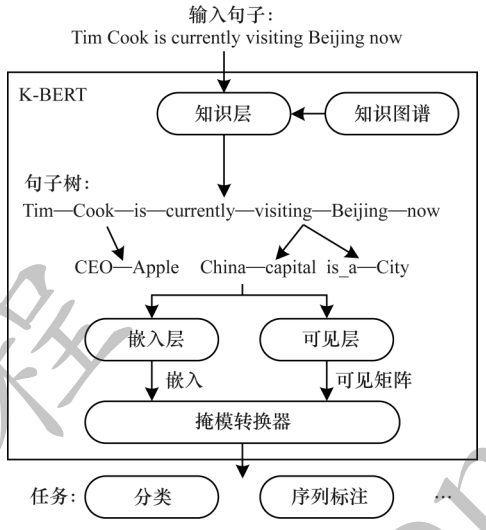

图2　K-BERT总体架构

Fig.2　The architecture of K-BERT

3.2.2　K-BERT实现过程

知识层完成知识注入和句子树转换，输入原始句子$s=\{w_0,w_1,\cdots,w_n\}$和知识图谱K，输出句子树$t=\{w_0,w_1,\cdots,w_i\{(r_{i0},w_{i0}),\cdots,(r_{ik},w_{ik})\},\cdots,w_n\}$，整个过程分为知识查询和知识注入2个步骤：

1)知识查询选中输入句子中涉及的所有实体并查询它们在KGs中对应的三元组E，可以表示为：

$$E=\mathrm{K\_Query}(s,K) \tag{5}$$

其中：$E=\{(w_i,r_{i0},w_{i0}),\cdots,(w_i,r_{ik},w_{ik})\}$是相应的三元组集合。

2)知识注入将查询到的三元组注入句子s，将E中三元组插入到相应的位置，并生成句子树t，可以表示为：

$$t=\mathrm{K\_Inject}(s,E) \tag{6}$$

嵌入层的作用是将输入的句子树转换为一种嵌入表示并保留其结构化信息，该嵌入表示是标记嵌入、位置嵌入、段嵌入等3个部分之和。

可见层是K-BERT和BERT之间最大的区别，K-BERT通过使用一个可见矩阵$\boldsymbol{M}$，限制每个标记的可见区域，从而规避过多知识注入引发句意改变的风险。

掩模转换器是掩模自注意块的层层堆叠，K-BERT提出的掩模自注意机制是自注意机制的延伸，如下所示：

$$Q^{i+1},K^{i+1},V^{i+1}=h^iW_q,h^iW_k,h^iW_v \tag{7}$$

$$S^{i+1}=\mathrm{softmax}\left(\frac{Q^{i+1}K^{i+1}+\boldsymbol{M}}{\sqrt{d_k}}\right) \tag{8}$$

$$h^{i+1}=S^{i+1}V^{i+1} \tag{9}$$

其中：W_q、W_k和W_v是可训练模型参数；h^i是第i个掩模自注意块的隐藏状态；d_k是比例因子。直观上，如果W_k对W_j不可见，则M_{jk}将把注意值S_{jk}^{i+1}掩盖为0，这意味着W_k对W_j的隐藏状态没有贡献。

3.2.3 K-BERT评估

将特定领域的数据集分为3个部分，分别用于模型的微调、选择和测试，各模型的检验结果如表2、表3所示。

表2 Google BERT模型的预训练表现

Table 2 Pre-training performance of Google BERT model %

模型	Finance_Q&A			Law_Q&A			Finance_NAR			Medicine_NER		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
Google BERT	81.9	86.0	83.9	83.1	90.1	86.4	84.8	87.4	86.1	91.9	93.1	92.5
K-BERT(How Net)	83.3	84.4	83.9	83.7	91.2	87.3	86.3	89.0	87.6	93.2	93.3	93.3
K-BERT(CN-DBpedia)	81.5	88.6	84.9	82.1	93.8	87.5	86.1	88.7	87.4	93.9	93.8	93.8
K-BERT(MedicalKGt)	—	—	—	—	—	—	—	—	—	94.0	94.4	94.2

表3 改进BERT模型的预训练表现

Table 3 Pre-training performance of the improved BERT model %

模型	Finance_Q&A			Law_Q&A			Finance_NAR			Medicine_NER		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
改进BERT	82.1	86.5	84.2	83.2	91.7	87.2	84.9	87.4	86.1	91.8	93.5	92.7
K-BERT(How Net)	82.8	85.8	84.3	83.0	92.4	87.5	86.3	88.5	87.3	93.5	93.8	93.7
K-BERT(CN-DBpedia)	81.9	87.1	84.4	83.1	92.6	87.6	86.3	88.6	87.4	93.9	94.3	94.1
K-BERT(MedicalKGt)	—	—	—	—	—	—	—	—	—	94.1	94.3	94.2

从表2、表3可以看出，在特定领域的任务(包括金融、法律和医学)中，K-BERT的表现明显优于BERT，尤其在医学领域效果提升显著，这表明K-BERT是解决需要专家知识驱动问题的较佳方案。

上文详细阐述了K-BERT将三元组作为领域知识注入到句子中，并引入可见矩阵限制知识影响以解决噪声问题的过程。

3.2.4 其他领域知识融合的预训练语言模型

上文所述方法大多在注入知识时更新预训练模型的原始参数，因此，在注入多种知识时可能会遭遇灾难性遗忘问题。为此，文献[52]提出一种K-Adapter模型，其可以使BERT自适应地与知识相融合：先针对不同的预训练任务对不同的adapter进行独立训练，然后在针对具体的下游任务进行微调时，采用不同的adapter有针对性地加入特征以增强其效果。这样不仅保持了预训练模型的原始参数不变，还能进行持续的知识灌输。

3.3 常识知识融合的预训练语言模型

常识知识对某些NLP任务至关重要，其中，大规模常识知识如何促进语言的理解和生成是关键，特别是对话系统，原因是会话交互是一种语义活动[53]，如果模型能够访问和充分利用大规模的常识知识，就能更好地理解对话，从而做出更恰当的响应。目前已有研究人员在会话生成中通过引入外部知识来增强对话生成，并取得了一定成果，但在会话生成方面还有一些问题未能解决：

1)一个实体词通常可以指不同的概念，即一个实体有多种含义，但在特定的语境中只涉及一个特定的概念，如果不考虑这一点，一些预先获取的知识事实可能与上下文无关。

2)即使只考虑一个特定的实体意义，相关的知识事实也可能涵盖各种目标主题，但是，其中一些主题并不有助于产生对话。

3)以往方法在知识整合与对话生成方面存在不足，包括整合的方式以及知识的类型。

本节介绍一种全新的常识知识感知的对话生成模型ConKADI[54]，该模型可以很好地解决以上问题。

3.3.1 ConKADI总体架构

ConKADI总体架构如图3所示，知识事实集F由知识检索器(Knowledge Retriever)根据查询信息X进行检索，上下文编码器(Context Encoder)将语句总结为上下文表示形式。虚构事实识别器(Felicitous Fact Recognize)计算出合适的事实概率分布z除以F，用于初始化解码器并指导其生成。三元组知识解码器(Triple Knowledge Decoder)可以生成词汇、知识实体词和复制词，并采用灵活模式融合(Flexible Mode Fusion)。

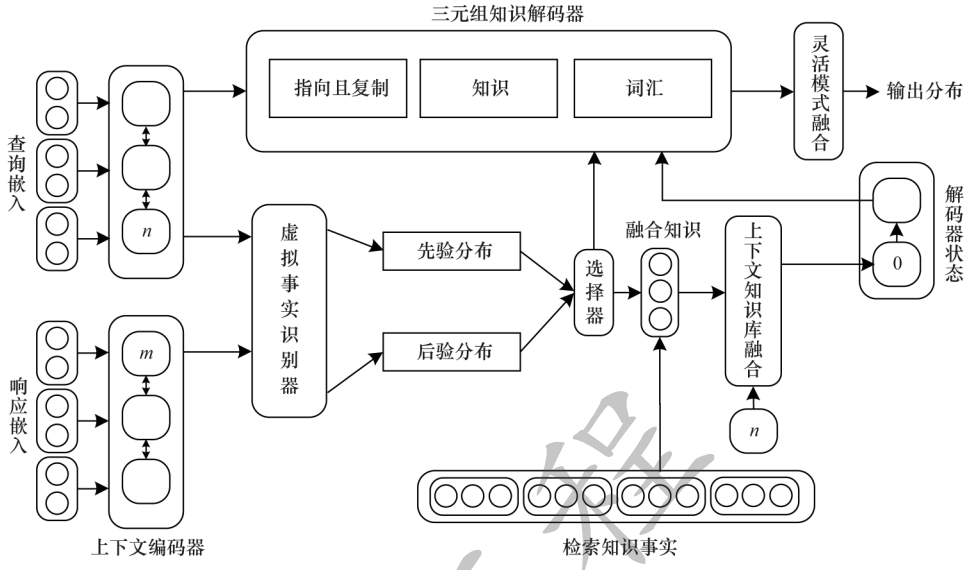

图3　ConKADI总体架构

Fig.3　The architecture of ConKADI

各模块具体如下：

1）知识检索器。给定一个查询消息X，如果一个词$x_i \in X$被识别为一个实体词，并且可以匹配到知识图谱G中的顶点e_{src}，若其他顶点$e_{\text{tgt}} \in \text{Neighbour}(e_{\text{src}})$，将检索的相应方向关系$r$作为候选事实$f$。其中，$e_{\text{src}}$、$e_{\text{tgt}}$称为源、目标实体。

2）上下文编码器。该编码器是一个双向GRU网络，其读取X或Y并输出上下文状态序列。

3）虚构事实识别器。某些初步检索到的知识事实在对话上下文中可能不合适，虚构事实识别器旨在检测与对话上下文高度一致的事实，其读取上下文信息，然后在F上输出概率分布$z \in \mathbb{R}^{l\times 1}$，第$i$维值$z[i]$表示$f_i$的权重。在训练阶段，将高质量的人为反应$Y$作为后验知识。因此，在训练中采用后验$z_{\text{post}}$，在推理中采用先验$z_{\text{prior}}$：

$$z_{\text{post}} = \eta\left(\phi\left(\boldsymbol{F}\cdot W_{ft}\right)\cdot\phi\left(\left[\boldsymbol{h}_n^{x\,\mathrm{T}};\ \boldsymbol{h}_m^{y\,\mathrm{T}}\right]\cdot W_{\text{post}}\right)\right)^{\mathrm{T}} \tag{10}$$

$$z_{\text{prior}} = \eta\left(\phi\left(\boldsymbol{F}\cdot W_{ft}\right)\cdot\phi\left(\boldsymbol{h}_n^{x\,\mathrm{T}}\cdot W_{\text{prior}}\right)\right)^{\mathrm{T}} \tag{11}$$

其中：$\boldsymbol{F}$是检索到的事实的嵌入矩阵；W_{ft}、W_{post}和W_{prior}是可训练参数；η是softmax激活函数；ϕ是tanh激活函数。Kullback- Leibler散度（KLD）函数用来缩小2个分布之间的差距：

$$\ell_k = \text{KLD}\left(z_{\text{post}}, z_{\text{prior}}\right) \tag{12}$$

基于知识融合和上下文检索，对解码器进行初始化以增强其对背景知识的理解：

$$\boldsymbol{h}_0^{y\,\mathrm{T}} = \tanh\left(\left[\boldsymbol{h}_n^{x\,\mathrm{T}};\ \boldsymbol{f}_z^{\,\mathrm{T}}\right]\cdot W_{\text{init}}\right) \tag{13}$$

采用词袋丢失以确保上下文$\boldsymbol{h}_n^x$与知识融合$\boldsymbol{f}_z$输入的准确性，通过构造一个0-1指标向量$\boldsymbol{I}_f \in \mathbb{R}^{l\times 1}$来监督$z_{\text{post}}$的训练，目的是通过以下方式使$\ell_f$达到最小化：

$$-\frac{\sum_{y_b\in B}\log_a p_b\left(y_b \mid \boldsymbol{h}_n^x, \boldsymbol{f}_z\right)}{|B|} - \frac{I_f^{\mathrm{T}}\cdot\log_a\left(z_{\text{post}}\right)}{|I_f|} \tag{14}$$

其中：B是Y的词袋；p_b是被softmax函数激活的双层MLP_{bow}，由它输出词汇V的概率分布。

4）三元组知识解码器。该解码器是另一个GRU网络，在每个时间步可以生成词汇、知识实体词和复制词中的一种。词汇的概率分布为：

$$\boldsymbol{p}_{w,t}^{\mathrm{T}} = \eta\left(\text{elu}\left(\left[\boldsymbol{h}_t^{y\,\mathrm{T}};\ \boldsymbol{u}_{t-1}^{\mathrm{T}};\ \boldsymbol{c}_t^{\mathrm{T}}\right]\cdot W_{v1}\right)\cdot W_{v2}\right) \tag{15}$$

其中：elu是非线性层激活函数；W_{v1}、W_{v2}是可训练参数；$\boldsymbol{h}_t^y$是ConKADI首先更新的内部状态。

知识实体词的概率分布为：

$$z_{d,t} = \eta\left(\phi\left(\boldsymbol{F}\cdot W_{fd}\right)\cdot\phi\left(\left[\boldsymbol{h}_t^{y\,\mathrm{T}};\ \boldsymbol{u}_{t-1}^{\mathrm{T}}\right]\cdot W_d\right)^{\mathrm{T}}\right) \tag{16}$$

$$\gamma_t = \text{sigmoid}\left(\left[\boldsymbol{h}_t^{y\,\mathrm{T}};\ \boldsymbol{u}_t^{\mathrm{T}};\ \boldsymbol{c}_t^{\mathrm{T}}\right]\cdot W_{\text{gate}}\right) \in \mathbb{R} \tag{17}$$

$$p_{k,t} = \gamma_t \times z + \left(1.0-\gamma_t\right)\times z_d \tag{18}$$

其中：z是静态全局分布；z_d是动态分布。解码器从X中指定一个单词x，然后复制x，查询信息x对应的概率分布为：

$$p_{c,t} = \eta\left(\phi\left(H^x\cdot W_{cs}\right)\cdot\phi\left(\boldsymbol{u}_t^{c\,\mathrm{T}}\cdot W_{ct}\right)^{\mathrm{T}}\right) \tag{19}$$

$$\boldsymbol{u}_t^{c\,\mathrm{T}} = \left[\boldsymbol{h}_t^{y\,\mathrm{T}};\ \boldsymbol{u}_{t-1}^{\mathrm{T}};\ \boldsymbol{c}_t^{\mathrm{T}}\right] \tag{20}$$

随后通过最小化ℓ_n来优化融合输出分布$p\left(Y \mid X, \boldsymbol{F}\right)$，具体为：

$$-\sum_t \lambda_t \log_a p_{\text{out},t}\left(y_t \mid y_{t-1;1}, X, \boldsymbol{F}\right) + \frac{\ell_m}{2} \tag{21}$$

训练目标是通过最小化如下ℓ来训练ConKADI：

$$\ell = \ell_n + \ell_k + \ell_f \tag{22}$$

3.3.2　ConKADI评估

为了研究不同语言之间的泛化性，实验采用公开的英语Reddit数据集、中文微博数据集，基线选取

S2S、ATS2S、GenDS和CCM等模型，从知识利用率、基于嵌入的相关性、基于重叠的相关性、多样性和信息量这5个方面进行评价。为了验证模型的综合性能，先按平均指标（AVG）计算7个基线的平均得分，然后计算算术平均得分如下：

$$R_{\mathrm{a}}=\frac{1}{5}\sum_{A_i}\left(\frac{1}{|A_i|}\sum_{m\in A_i}\frac{m_j}{m_{j,\mathrm{AVG}}}\right) \tag{23}$$

计算几何平均得分如下：

$$R_{\mathrm{g}}=\left(\prod_{A_i}\left(\prod_{m_j\in A_i}\frac{m_j}{m_{j,\mathrm{AVG}}}\right)^{\frac{1}{|A_i|}}\right)^{\frac{1}{5}} \tag{24}$$

实验结果如表4、表5所示，从表4、表5可以看出：在相对得分方面，ConKADI的整体表现优于基线模型，与最先进的CCM方法相比，ConKADI在中文数据集上的得分分别提高了153%（算术）、95%（几何），在英文数据集上分别提高了48%（算术）、25%（几何）；在知识利用率方面，GenDS、CCM、ConKADI等3种知识感知模型的表现显著优于其他模型，而与GenDS、CCM相比，ConKADI对知识的利用率更高，并且更能找出准确的知识；在多样性和信息性方面，ConKADI性能提升显著归因于上下文-知识融合；在相关性方面，由于数据集存在内在差异，因此ConKADI在中文数据集上的整体性能最好，但在英文数据集上其性能并不理想。

表4 各模型在中文微博数据集上的性能对比结果

Table 4 Performance comparison results of each model on Chinese microblog dataset

数据集	实体得分			嵌入		BLUE-2	BLUE-3	Distinct-1	Distinct-2	Entropy	R-Sore	
	E_{match}	E_{use}	E_{recall}/%	$\mathrm{Emb}_{\mathrm{avg}}$	$\mathrm{Emb}_{\mathrm{ex}}$						R_{a}	R_{g}
S2S	0.33	0.58	13	0.770	0.500	2.24	0.80	0.21	1.04	6.09	0.78	0.75
ATS2S	0.33	0.59	12	0.767	0.513	1.93	0.69	0.27	1.23	5.99	0.77	0.75
$\mathrm{ATS2S}_{\mathrm{MMI}}$	0.40	0.74	15	0.773	0.528	4.01	1.61	0.75	3.91	7.49	1.24	1.21
$\mathrm{ATS2S}_{\mathrm{DD1.5}}$	0.35	0.62	13	0.780	0.542	2.14	0.86	1.03	4.86	7.62	1.16	1.10
Copy	0.33	0.68	13	0.786	0.501	2.28	0.84	0.59	2.18	6.13	0.92	0.91
GenDS	0.75	0.84	26	0.789	0.524	2.09	0.73	0.30	1.66	5.89	0.94	0.91
CCM	0.99	1.09	28	0.786	0.544	3.26	1.20	0.48	2.59	6.16	1.18	1.15
AVG	0.49	0.74	17	0.799	0.522	2.56	0.96	0.52	2.50	6.48	1.00	1.00
ConKADI	1.48	2.08	38	0.846	0.577	5.06	1.59	3.26	23.93	9.04	2.98	2.24
$\mathrm{ConKADI}_{\mathrm{-cp}}$	1.60	1.89	38	0.833	0.567	5.00	1.52	2.34	18.29	8.75	2.55	2.08

表5 各模型在英文Reddit数据集上的性能对比结果

Table 5 Performance comparison results of each model on English Reddit dataset

数据集	实体得分			嵌入		BLUE-2	BLUE-3	Distinct-1	Distinct-2	Entropy	R-Sore	
	E_{match}	E_{use}	E_{recall}/%	$\mathrm{Emb}_{\mathrm{avg}}$	$\mathrm{Emb}_{\mathrm{ex}}$						R_{a}	R_{g}
S2S	0.41	0.52	4	0.868	0.837	4.81	1.89	0.38	1.77	7.59	0.82	0.78
ATS2S	0.44	0.59	5	0.863	0.831	4.50	1.81	0.82	3.44	7.62	0.92	0.91
$\mathrm{ATS2S}_{\mathrm{MMI}}$	0.45	0.65	6	0.858	0.825	4.95	2.13	0.75	3.22	7.62	0.95	0.94
$\mathrm{ATS2S}_{\mathrm{DD1.5}}$	0.31	0.43	4	0.830	0.784	1.70	0.75	0.97	3.50	7.47	0.77	0.72
Copy	0.13	0.67	9	0.868	0.841	5.43	2.26	1.73	8.33	7.87	1.19	1.09
GenDS	1.13	1.26	13	0.876	0.851	4.68	1.79	0.74	3.97	7.73	1.14	1.10
CCM	1.08	1.33	11	0.871	0.841	5.18	2.01	1.05	5.29	7.73	1.21	1.18
AVG	0.55	0.77	7	0.860	0.829	4.40	1.79	0.94	4.32	7.69	1.00	1.00
ConKADI	1.24	1.98	14	0.867	0.852	3.53	1.27	2.77	18.78	8.50	1.76	1.46
$\mathrm{ConKADI}_{\mathrm{-cp}}$	1.41	1.73	13	0.865	0.855	3.09	1.07	2.29	16.70	8.68	1.63	1.37

综上，ConKADI模型提出的上下文知识融合和灵活模式融合可以促进知识整合，在开放发布的英文数据集和中文数据集中取得显著成果，但其在推进高质量知识整合方面仍存在一定的局限性。

3.3.3 其他常识知识融合的预训练语言模型

在其他常识知识融合的预训练语言模型中，常识知识感知的会话生成模型CCM[55]与现有单独使用知识三元组（实体）的模型不同，CCM将每个知识图视为一个整体，在图中编码结构化更强、连接更紧密的语义信息。具体来说，CCM通过从一个知识库中检索出相关的知识图谱，并采用静态的图注意机制对其进行编码，促进开放领域会话系统中的语言理解和生成，然后通过动态图注意机制在词生成过程中检索知识图和图中的每个知识三元组，以便更

好地完成生成任务。

3.4　语义知识融合的预训练语言模型

受BERT掩蔽策略的启发，百度提出一种新的知识增强语言表示模型ERNIE(baidu)，用来学习通过知识掩蔽策略增强的语言表示。ERNIE(baidu)的掩蔽策略如图4所示。

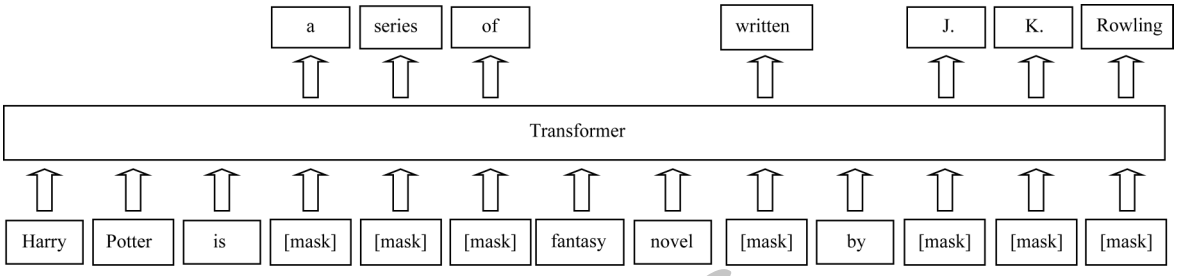

图4　ERNIE(baidu)的掩蔽策略

Fig.4　The masking strategy of ERNIE(baidu)

3.4.1　ERNIE(baidu)的Transfomer编码器

与BERT相同，ERNIE(baidu)使用多层Transfomer作为基础编码器，Transformer可以通过自注意机制获取句子中每个标记的上下文信息，并生成上下文嵌入序列。

对于汉语语料库，本文使用词块对汉语句子进行标记。对于给定的标记，其输入表示是由相应的令牌、段和位置嵌入的总和构成的，每个序列的第一个标记是特殊分类嵌入([CLS])。

3.4.2　ERNIE(baidu)的知识集成方法

知识集成的主要思路是使用先验知识增强预训练语言模型。ERNIE(baidu)提出一种多阶段知识掩蔽策略，将短语级和实体级知识整合到语言表示中，而非直接添加知识嵌入。句子的不同掩蔽级别如图5所示。

Sentence	Harry	Potter	is	a	series	of	fantasy	novel	written	by	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novel	[mask]	by	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novel	[mask]	by	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novel	[mask]	by	[mask]	[mask]	[mask]

图5　句子的不同掩蔽级别

Fig.5　Different masking levels of a sentence

第一阶段是基础级掩蔽，在训练过程中随机掩码15%的基本语言单元，并使用句子中其他基本语言单元作为输入，训练一个Transformer来预测掩码单元，由此得到的基本词表示很难对高级语义知识进行完全建模。

第二阶段是短语级掩蔽，同样使用基本语言单元作为训练输入，不同于上一阶段，这次选择随机掩蔽几个短语，并预测同一短语中的所有基本单元。在此阶段，短语信息被编码到词嵌入中。

第三阶段是实体级掩蔽，实体通常在句子中包含重要信息。在上一阶段，分析得到句子中的命名实体并将其掩蔽，随后进行预测并填补空缺的实体。

经过上述3个阶段的学习，能够得到经过丰富语义信息增强后的词表示。

3.4.3　ERNIE(baidu)评估

ERNIE(baidu)和BERT在自然语言推理、语义相似度、命名实体识别、情感分析和问答这5个中文自然语言处理任务中的性能表现如表6所示。

表6　ERNIE(baidu)和BERT在5个中文自然语言处理任务中的性能对比

Table 6　Performance comparison of ERNIE(baidu) and BERT in five Chinese NLP tasks　%

任务	指标	BERT		ERNIE(baidu)	
		dev	test	dev	test
XNLI	精度	78.1	77.2	79.9(+1.8)	78.4(+1.2)
LCQMC	精度	88.8	87.0	89.7(+0.9)	89.4(+0.4)
MSRA-NER	F1	94.0	92.6	95.0(+1.0)	93.8(+1.2)
ChnSentiCorp	精度	94.6	94.3	95.2(+0.6)	95.4(+1.1)
nlpcc-dbqa	mrr	94.7	94.6	95.0(+0.3)	95.1(+0.5)
	F1	80.7	80.8	82.3(+1.6)	82.7(+1.9)

从表6可以看出，通过对异构数据的知识整合和预训练，模型的语义表示能力显著增强，能够获得更好的语言表达。在上述5个中文自然语言处理任务中，ERNIE(baidu)的表现均明显优于BERT。

以上详细阐述了ERNIE(baidu)在BERT的基础上使用Transformer[56]作为特征抽取器，加入海量语料中的短语、实体等先验语义知识，从而建模真实

世界语义关系的过程。与BERT相比,ERNIE(baidu)仅对学习任务MLM稍作改进就取得了较好的成效。

3.4.4 其他语义知识融合的预训练语言模型

ERNIE(baidu)的改进模型ERNIE 2.0是基于持续学习的语义理解预训练框架,使用多任务学习增量式构建预训练任务,新增的语义任务包括预测句中词汇、重建和重排句子结构、判断句间逻辑关系等。采用语言学及语义知识注入的预训练语言模型还有面向中文理解的神经语境表征模型NEZHA[57],其在BERT的基础上进行的改进包括作为有效位置编码方案的函数相对位置编码、全词掩蔽策略、混合精度训练和训练模型的LAMB优化器。使用N-gram[58]表示增强的中文文本编码器的预训练模型ZEN[59],其基于BERT,在训练过程中考虑不同的字符组合,整合字符序列及其所包含的单词或短语的综合信息,相比于其他编码器使用更少的资源,却能在大多数任务上展现出更佳的性能。基于BERT的汉语全词掩蔽预训练语言模型BERT-WWM-Chinese[60],当被掩蔽的词块标记属于整词时,所有构成整词的词块标记都将被一并掩蔽,迫使模型在掩蔽语言模型(MLM)的预训练任务中恢复整个单词,而不仅仅恢复单词标记。上述模型均通过对BERT进行微小改进而使得表征能力和训练速度大幅提升。

3.5 专业知识融合的预训练语言模型

本节以预训练技术在生物医学领域的应用为代表,介绍将特定领域的专业知识注入到预训练语言模型的相关研究。

随着生物医学文档数量的快速增长,生物医学文本挖掘需求日益增大。BioBERT模型[61]是一种注入了生物医学领域知识的预训练语言模型,其基于BERT,致力于合理利用生物医学语料库,达到最佳的预期效果以帮助临床实践。BioBERT的预训练和微调过程如图6所示。

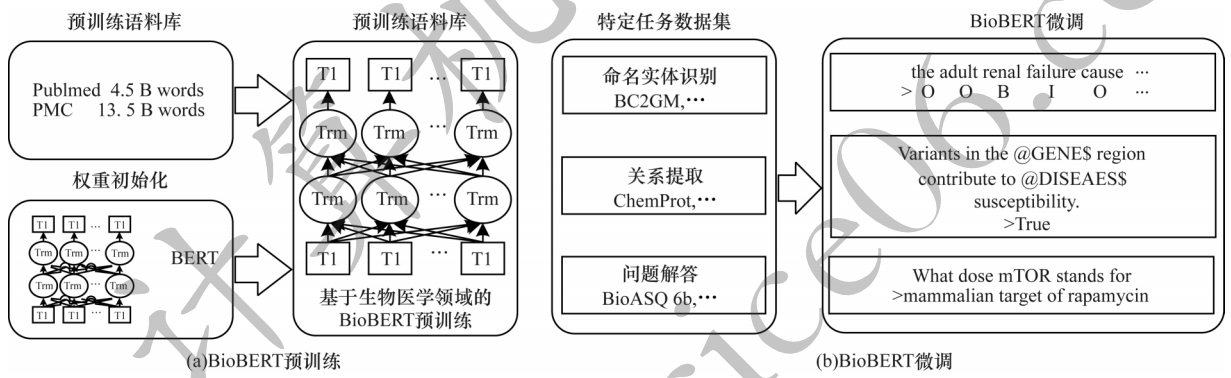

图6 BioBERT的预训练和微调流程

Fig.6 Pre-training and fine-tuning procedure of BioBERT

3.5.1 BioBERT预训练

首先使用BERT的权重初始化BioBERT,该BERT已在通用领域语料库(英语Wikipedia和BooksCorpus)上进行了预训练;然后对BioBERT进行生物医学领域语料库(PubMed摘要和PMC全文文章)的预训练。用于BioBERT预训练的文本语料库如表7所示,经过测试的文本语料库组合如表8所示。

表7 BioBERT使用的文本语料库列表

Table 7 List of text corpora used by BioBERT

语料库	词量/B	领域
English Wikipedia	2.5	公共领域
BooksCorpus	0.8	公共领域
PubMed Abstracts	4.5	生物医学领域
PMC Full-text atticles	13.5	生物医学领域

表8 文本语料库的不同组合

Table 8 Different combinations of text corpora

模型	语料组合
BERT	Wiki+Books
BioBERT(+PubMed)	Wiki+Books+PubMed
BioBERT(+PMC)	Wiki+Books+PMC
BioBERT(+PubMed+PMC)	Wiki+Books+PubMed+PMC

3.5.2 BioBERT微调

本文针对命名实体识别(NER)、关系提取(RE)和问题解答(QA)这3个具有代表性的生物医学文本挖掘任务对BioBERT进行调整。

NER是最基本的生物医学文本挖掘任务之一,涉及识别生物医学语料库中的众多领域特定专有名词,在预训练和微调过程中,BioBERT直接学习词块嵌入,使用实体级精度、召回率和F1值作为NER的评价指标。

QA是一种用自然语言对相关段落提出的问题进行回答的任务,为了对BioBERT的QA进行微调,本文采用SQuAD数据集的BERT架构。由于

BioASQ的格式与SQuAD相似，BioBERT使用了BioASQ事实数据集，利用单个输出层计算答案短语开始/结束位置的令牌级别概率。然而，约30%的BioASQ问题在抽取的QA设置中无法回答，因为确切的答案没有出现在给定的段落中，所以将无法回答问题的样本从训练集中剔除。使用strict accuracy、lenient accuracy和mrr（mean reciprocal rank）作为QA的评价指标。mrr得分的计算公式为：

$$M_{\text{mrr}}=\frac{1}{|Q|}\left|\sum_{i=1}^{|Q|}\frac{1}{R_{\text{rank}}^{i}}\right| \tag{25}$$

3.5.3　BioBERT评估

本次实验使用较多生物医学NLP研究人员常用的数据集，包括NCBI疾病、BC5CDR、BC4CHEMD等，本文将BERT和BioBERT与当前最新模型进行比较，并报告其得分。

NER结果如表9所示，从表9可以看出：仅对通用域语料库进行预训练的BERT具有有效性，但是BERT的微观平均F1得分比最新模型低（2.01）；另一方面，在所有数据集上，BioBERT的得分均高于BERT，在9个数据集中，有6个数据集上的BioBERT性能优于最新模型。RE结果如表10所示，从表10可以看出，BERT的性能优于CHEMPROT数据集上的最新模型，这证明了其在RE中的有效性。从平均水平（微型）来看，BioBERT v1.0（+PubMed）的F1得分高于最新模型。此外，BioBERT在2个生物医学数据集中均获得了最高的F1分数。QA结果如表11所示，从表11可以看出，BioBERT各项性能明显优于BERT和最新模型。

表9　生物医学命名实体识别测试结果

Table 9　Test results of biomedical named entity recognition　%

数据集	SOFT			BERT (Wiki+Books)			BioBERT v1.0 (+PubMed)			BioBERT v1.0 (+PMC)			BioBERT v1.0 (+PubMed+PMC)			BioBERT v1.1 (+PubMed)		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
NVBI disease	88.30	89.00	86.60	84.12	87.17	85.63	86.76	88.02	87.38	86.16	89.48	87.79	89.04	89.69	89.36	88.22	91.25	89.71
2010 i2b2/VA	87.44	86.25	86.84	84.04	84.08	84.06	85.37	85.64	85.51	85.55	85.72	85.64	87.50	85.44	86.46	86.93	86.53	86.73
BC5CDR（Disease）	89.61	83.09	86.23	81.97	82.48	82.41	85.80	86.60	86.20	84.67	85.87	85.27	85.86	87.27	86.56	86.47	87.84	87.15
BC5CDR (Drug/chem)	94.26	92.38	93.31	90.94	91.38	91.16	92.52	92.76	92.64	92.46	92.63	92.54	93.27	93.61	93.44	93.68	93.26	93.47
BC4CHEMD	92.29	90.01	91.14	91.19	88.92	90.04	91.77	90.77	91.26	91.65	90.30	90.97	92.23	90.61	91.41	92.80	91.92	92.36
BC2GM	81.81	81.57	81.69	81.17	82.42	81.79	81.72	83.38	82.54	82.86	84.21	83.53	85.16	83.65	84.40	84.32	85.12	84.72
JNLPBA	74.43	83.22	78.58	69.57	81.20	74.94	71.11	83.11	76.65	71.17	82.76	76.53	72.68	83.21	77.59	72.24	83.56	77.49
LINNAEUS	92.80	94.29	93.54	91.17	84.30	87.60	91.83	84.72	88.13	91.62	85.48	88.45	93.84	86.11	89.81	90.77	85.83	88.24
Species-800	74.34	75.96	74.58	69.35	74.05	71.63	70.60	75.75	73.08	71.54	74.71	73.06	72.84	77.97	75.31	72.80	75.36	74.06

表10　生物医学关系提取测试结果

Table 10　Test results of biomedical relation extraction　%

数据集	SOFT			BERT（Wiki+Books）			BioBERT v1.0（+PubMed）			BioBERT v1.1（+PubMed）		
	P	*R*	F1	*P*	*R*	F1	*P*	*R*	F1	*P*	*R*	F1
GAD	79.21	89.25	83.93	74.28	85.11	79.29	76.43	87.65	81.61	77.32	82.68	79.83
EU-ADR	76.43	98.01	85.34	75.45	96.55	84.62	78.04	93.86	84.44	77.86	83.55	89.74
CHEMPROT	74.80	56.00	64.10	76.02	71.60	73.74	76.05	74.33	75.18	77.02	75.90	76.46

表11　生物医学问题回答测试结果

Table 11　Test results of biomedical question answering　%

数据集	SOFT			BERT（Wiki+Books）			BioBERT v1.0（+PubMed+PMC）			BioBERT v1.1（+PubMed）		
	S	*L*	*M*	*S*	*L*	*M*	*S*	*L*	*M*	*S*	*L*	*M*
BioASQ 4b	20.01	28.81	23.52	27.33	44.72	33.77	28.57	47.82	35.17	27.95	44.10	34.72
BioASQ 5b	41.33	56.67	47.24	39.33	52.67	44.27	44.00	56.67	49.38	46.00	60.00	51.64
BioASQ 6b	24.22	37.89	27.84	33.54	51.55	40.88	40.37	57.77	47.48	42.86	57.77	48.43

以上主要介绍用于生物医学文本挖掘的预训练语言表示模型BioBERT，实验结果证明，在生物医学语料库上对BERT引入特定领域知识，对于将其应用于生物医学领域至关重要。只需对特定任务进行微小的架构修改，BioBERT就可以在生物医学文本挖掘任务（如NER、RE和QA等）中取得优异成果。

3.5.4 其他专业知识融合的预训练语言模型

基于BERT语言模型的专利分类模型PatentBERT[62]通过使用带有单词嵌入的CNN方法，集中对一个预训练BERT模型做适当微调，并将其应用于专利分类这个多标签分类任务。面向科学文本的预训练语言模型SciBERT[63]利用大型多领域科学出版物语料库进行无监督的预训练，以提高下游科学自然语言处理任务的效果。

3.6 模型性能总结对比

前文根据外部知识的分类，分别对语言学、语义、事实、常识、领域知识等多种类型知识注入的预训练语言模型进行详细阐述，实验结果证明，在知识驱动的NLP下游任务中使用知识感知的预训练语言模型，有助于提升工作效率。

通过对以上知识感知的预训练语言模型在执行常识和推理任务时所需的信息资源进行分析，本文评估并归纳不同预训练语言模型的综合性能，结果如表12所示。

表12 知识感知的预训练语言模型性能对比结果

Table 12 Performance comparison results of knowledge-aware pre-trained language models

预训练模型	特征提取器	训练任务	数据集/语料库	优点	缺点	适用场景
ERNIE (THU)	Transformer	MLM+NSP+dEA	WikiEn+Wikidata	在知识驱动的实体键入和关系分类任务中表现优异	依赖于NER提取的准确度，模型复杂度过高	自然语言推理、文本分类、问答任务
KG-BERT	Transformer	TC+LP+RP+AV	WordNet（WN11、WN18RR）+Freebase（FB15K、FB15K-237）+FB13+UMLS	在三元组分类、链接预测和关系预测任务中表现优异	尚未应用于语言理解任务	三元组分类、链接预测、关系预测任务
KnowBERT	Transformer	MLM+NSP+EL	WikiEn+WordNet/Wiki	擅长处理歧义问题，在实体键入和关系抽取任务中表现优异	不利于最大化AIDA的链接性能	关系抽取、实体键入和词义消歧任务
KEPLER	Transformer	MLM+KE	WikiEn+Wikidata /WordNet	能将事实知识有效整合到PLMs中，并能有效学习知识图嵌入	实体提及不突出，没有足够的训练任务以回顾事实性知识	关系分类、实体键入任务
K-BERT	Transformer	MLM+NSP	WikiZh + WebtextZh + CN-DBpedia + HowNet + MedicalKG	在特定领域的任务中表现优异	K-Query还需改进，未将该方法扩展到其他LR模型上	特定领域任务
K-Adapter	Transformer	RC+DRP	LAMA+LAMA-UHN	在实体键入、问答和关系分类任务中展现了最佳性能	在其他NLP任务中的性能还需提升	实体键入、关系分类、问答任务
ConKADI	RNN	FFM+TKD	English Reddit+Chinese Weibo	在中文知识感知对话生成任务中表现优异	在英文数据集上的性能并不理想	中文数据集上的对话生成任务
ERNIE (baidu)	Transformer	DLM+NSP+MLM	Chinese Wikepedia+Baidu Baike+Baidu News+Baidu Tieba	拥有出色的模型语义表示能力和对实体概念的学习推理能力	在除中文之外的其他语言任务中表现欠佳	自然语言推理、语义相似度、命名实体识别、情感分析、完形填空测验、问答任务
ERNIE 2.0	Transformer	MRC+NLI+SA+SS+QA	Encyclopedia+BookCorpus+News+Dialog+IR Relevance Data+Discourse	拥有出色的通用语义表示能力	持续学习的领域范围有待扩充	一般场景可直接使用，对性能要求严格的场景需先进行优化
NEZHA	Transformer	MLM+NSP	Chinese Wikipedia+Baidu Baike+Chinese News	在中文自然语言理解任务中表现优异	仅适用于有限场景，其他语言中的任务有待开发	命名实体识别、句子匹配、中文情感分类、跨语言自然语言推理任务
ZEN	Transformer	CWS+POS+NER+DC+SA+SPM+NLI	MSR+CBT5+MSRA+THUCNews+ChnSentiCorp6+LCQMC+XNLI	在中文自然语言处理的token级任务和序列级任务中效果显著	在句子级任务中效果不显著	中文分词、词性标注、命名实体识别、文本分类、情感分析、句子匹配、自然语言推理任务
BERT-WWM-Chinese	双向Transformer	MLM+MRC+NLI+SC+SPM+DC	ChnSentiCorp+BQCorpus+LCQMC	在中文文本处理任务中表现优异	在长序列任务中效果不显著	机器阅读理解、自然语言推理、情感分析、句子匹配、文本分类任务
BioBERT	双向Transformer	NER+RE+QA	Wikipedia+BooksCorpus+PubMed+PMC	在生物医学文本挖掘任务中性能明显优于BERT	存在专有受用范围，不普适	生物医学命名实体识别、生物医学关系抽取、生物医学问答任务
PatentBERT	Transformer	EPO+WIPO	IPC+Title+Abstract	在专利分类领域表现优异	存在专有受用范围，不普适	分类任务
SciBERT	Transformer	NER+PICO+CLS+REL+DEP	BASEVOCAB+SCIVOCAB	在科学文本方面表现优于BERT，在许多下游任务中实现了新的SOTA	暂时还无法在多领域使用单一的汇总模型	序列标注、句子分类、依存关系解析任务

4　研究前景展望

目前，知识感知的预训练语言模型在NLP任务中得到广泛应用并获得了良好收益，但是，一些不足之处也逐渐显现，其发展主要面临如下挑战：

1）预训练代价大。预训练模型的规模呈现逐步扩大的趋势，虽然其能力越来越强，但模型的参数也越来越多，因此，对计算设备的算力提出了更高的要求。预训练语言模型如何使用较小的代价来高质量地完成NLP任务是其当前面临的问题之一。

2）深层结构化语义分析存在明显的性能不足。当前大多数引入了语义知识的预训练语言模型仍难以抵御文本对抗攻击，当面向非规范和专用领域文本时，分析精度大幅降低，说明模型未能完全学习语言的真正寓意。仅在通用语料库上进行训练的模型往往经验不足，无法较好完成常识和推理任务，不符合人类使用语言的交际意图。

3）形式化知识在语言模型中存在明显构成缺失。该问题导致在问答任务中无法基于常识给出正确的回答，不足以完成细粒度的关系抽取，命名实体识别精度不高以及在知识图谱的知识系统中缺少动作、行为、状态以及事件逻辑关系的形式化描写。

4）跨模态的语言理解存在明显融通局限。跨模态关系因缺乏深层结构化语义分析和世界知识导致推理能力较弱，造成语言理解形意相离。

与此同时，知识感知的预训练语言模型已在现阶段部分知识驱动的NLP下游任务中崭露头角。预训练语言模型未来可能的研究方向如下：

1）降低预训练语言模型的训练成本。当前已有研究人员致力于在损失较小精度的前提下探索更小、更快的预训练语言模型，当前提供的思路包含对模型采用压缩和知识蒸馏[64]等方法，如百度公司发布的ERNIE-Tiny，将ERNIE 2.0 Base模型进行压缩，取得了4.3倍的预测提速。此外，研究新的高效的模型架构也将对降低训练消耗起到积极作用。

2）从海量的非结构化文本而非infobox中搜集系统的非结构化知识，构建新型多元知识图谱增强的预训练语言模型。在实体的静态事实描述较为充分的基础上，进一步完善事件的动态描述以及事件之间逻辑关系的构建。

3）未来可向大数据与富知识共同驱动的方向进行研究，并通过与图像等跨模态信息进行交互，提升以自然语言为核心的中文语义理解能力，在实践中探索经验主义与理性主义相结合的自然语言处理新范式。

5　结束语

本文阐述预训练技术的发展历程，分析其中经典的预训练语言模型，如BERT及其改进模型，归纳近年来预训练技术在NLP领域的应用现状和趋势。对现有预训练语言模型在NLP特定下游任务中的表现进行评估，结果表明，外部知识的引入对预训练语言模型在多种NLP下游任务中的性能提升起到积极作用。知识感知的预训练语言模型或可在更多特殊的应用情境中取得优异表现，逐步探索构建低成本、非结构化和可实现跨模态交互的预训练语言模型将是下一步的研究方向。

参考文献

[1] LIN Y，LIU Z，SUN M，et al. Learning entity and relation embeddings for knowledge graph completion [C]// Proceedings of AAAI Conference on Artificial Intelligence. New York，USA：AAAI Press，2015：134-145.

[2] 段丹丹，唐加山，温勇，等. 基于BERT模型的中文短文本分类算法[J]. 计算机工程，2021，47(1)：79-86.
DUAN D D，TANG J S，WEN Y，et al. Chinese short text classification algorithm based on BERT model [J]. Computer Engineering，2021，47(1)：79-86. (in Chinese)

[3] SOON W M，NG H T，LIM D C Y. A machine learning approach to coreference resolution of noun phrases [J]. Computational Linguistics，2001，27(4)：521-544.

[4] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving language understanding by generative pre-training [EB/OL]. [2021-01-05]. https：//www. cs. ubc. ca/~amuham01/LING530/papers/radford2018improving. pdf.

[5] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2021-01-05]. https：//aclanthology. org/N19-1423. pdf.

[6] QIU X，SUN T，XU Y，et al. Pre-trained models for natural language processing：a survey [EB/OL]. [2021-01-05]. https：//arxiv. org/pdf/2003. 08271v2. pdf.

[7] HOWARD J，RUDER S. Universal language model fine-tuning for text classification [EB/OL]. [2021-01-05]. https：//aclanthology. org/P18-1031. pdf.

[8] ROSSET C，XIONG C，PHAN M，et al. Knowledge-aware language model pretraining [EB/OL]. [2021-01-05]. https：//openreview. net/pdf？ id=OAdGsaptOXy.

[9] LIU H C，YOU J X，LI Z W，et al. Fuzzy petrinets for knowledge representation and reasoning：a literature review [J]. Engineering Applications of Artificial Intelligence，2017(60)：45-56.

[10] GUO S，WANG Q，WANG B，et al. SSE：semantically smooth embedding for knowledge graphs [J]. IEEE Transactions on Knowledge and Data Engineering，2017(29)：884-897.

[11] BORDES A，USUNIER N，GARCIA-DURAN A，et al. Translating embeddings for modeling multi-relational data [EB/OL]. [2021-01-05]. https：//proceedings. neurips. cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper. pdf.

[12] DING Y X, WU R, ZHANG X. Ontology-based knowledge representation for malware individuals and families [J]. Computers & Security, 2019(87): 101574.

[13] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261.
LIU Z Y, SUN M S, LIN Y K, et al. Knowledge representation learning: a review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261. (in Chinese)

[14] 卢晨阳,康雁,杨成荣,等. 基于语义结构的迁移学习文本特征对齐算法[J]. 计算机工程,2019,45(5):116-121.
LU C Y, KANG Y, YANG C R, et al. Text feature alignment algorithm for transfer learning based on semantic structure[J]. Computer Engineering, 2019, 45(5): 116-121. (in Chinese)

[15] WAN J, HUANG X. KaLM at SemEval-2020 task 4: knowledge-aware language models for comprehension and generation[EB/OL]. [2021-01-05]. https://arxiv. org/pdf/2005. 11768v2. pdf.

[16] SHI Y, ZHANG W Q, CAI M, et al. Efficient one-pass decoding with NNLM for speech recognition[J]. IEEE Signal Processing Letters, 2014, 21(4): 377-381.

[17] GOLDBERG Y, LEVY O. word2vec explained: deriving Mikolov et al. ' s negative-sampling word-embedding method[EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1402. 3722. pdf.

[18] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Washington D. C., USA: IEEE Press, 2014: 1532-1543.

[19] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[EB/OL]. [2021-01-05]. https://pdfs. semanticscholar. org/892e/53fe5cd39f037cb2a961499f42f3002595dd. pdf.

[20] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[EB/OL]. [2021-01-05]. https://aclanthology. org/P15-1150. pdf.

[21] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL]. [2021-01-05]. https://aclanthology. org/N18-1202. pdf.

[22] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[EB/OL]. [2021-01-05]. https://proceedings. neurips. cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper. pdf.

[23] KENTER T, BORISOV A, DE RIJKE M. Siamese CBOW: optimizing word embeddings for sentence representations [EB/OL]. [2021-01-05]. https://aclanthology. org/P16-1089. pdf.

[24] GUTHRIE D, ALLISON B, LIU W, et al. A closer look at skip-gram modeling[EB/OL]. [2021-01-05]. http://www. lrec-conf. org/proceedings/lrec2006/pdf/357_pdf. pdf.

[25] PENG H, LI J, SONG Y, et al. Incrementally learning the hierarchical softmax function for neural language models [EB/OL]. [2021-01-05]. http://home. cse. ust. hk/~yqsong/papers/2017-AAAI-Incremental. pdf.

[26] MCCANN B, BRADBURY J, XIONG C, et al. Learned in translation: contextualized word vectors[EB/OL]. [2021-01-05]. https://papers. nips. cc/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper. pdf.

[27] RESNIK P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language[J]. Journal of Artificial Intelligence Research, 1999, 11: 95-130.

[28] 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.

[29] LEVY O, GOLDBERG Y. Neural word embedding as implicit matrix factorization [J]. Advances in Neural Information Processing Systems, 2014, 27: 2177-2185.

[30] GREFF K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(10): 2222-2232.

[31] WU X, ZHANG T, ZANG L, et al. "Mask and infill": applying masked language model to sentiment transfer[EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1908. 08039. pdf.

[32] GHAZVININEJAD M, LEVY O, LIU Y, et al. Mask-predict: parallel decoding of conditional masked language models[EB/OL]. [2021-01-05]. https://aclanthology. org/D19-1633. pdf.

[33] SONG K, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation[EB/OL]. [2021-01-05]. https://www. microsoft. com/en-us/research/uploads/prod/2019/06/MASS-paper-updated-002. pdf.

[34] DONG L, YANG N, WANG W, et al. Unified language model pre-training for natural language understanding and generation[EB/OL]. [2021-01-05]. https://papers. nips. cc/paper/2019/file/c20bb2d9a50d5ac1f713f8b34d9aac5a-Paper. pdf.

[35] CHENG Y, FU S, TANG M, et al. Multi-task deep neural network enabled optical performance monitoring from directly detected PDM-QAM signals[J]. Optics Express, 2019, 27(13): 19062-19074.

[36] SUN Y, WANG S, LI Y, et al. ERNIE 2. 0: a continual pre-training framework for language understanding [C]// Proceedings of AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020: 8968-8975.

[37] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities[EB/OL]. [2021-01-05]. https://aclanthology. org/P19-1139. pdf.

[38] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[EB/OL]. [2021-01-05]. https://

research. fb. com/wp-content/uploads/2017/09/emnlp2017. pdf.
[39] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[EB/OL]. [2021-01-05]. https://aclanthology. org/N16-1030. pdf.
[40] LIU Y, OTT M, GOYAL N, et al. RoBERT: a robustly optimized bert pretraining approach[EB/OL]. [2021-01-05]. https://export. arxiv. org/pdf/1907. 11692.
[41] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT [EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1906. 08101v2. pdf.
[42] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [EB/OL]. [2021-01-05]. https://openreview. net/pdf? id=H1eA7AEtvS.
[43] JOSHI M, CHEN D, LIU Y, et al. SpanBERT: improving pre-training by representing and predicting spans[EB/OL]. [2021-01-05]. https://www. cs. princeton. edu/~danqic/papers/tacl2020. pdf.
[44] JIAO X, YIN Y, SHANG L, et al. TinyBERT: distilling BERT for natural language understanding[EB/OL]. [2021-01-05]. https://aclanthology. org/2020. findings-emnlp. 372. pdf.
[45] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.
[46] AGARWAL O, GE H, SHAKERI S, et al. Large scale knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training[EB/OL]. [2021-01-05]. https://aclanthology. org/2021. naacl-main. 278. pdf.
[47] YAO L, MAO C, LUO Y. KG-BERT: BERT for knowledge graph completion[EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1909. 03193. pdf.
[48] PETERS M E, NEUMANN M, LOGAN IV R L, et al. Knowledge enhanced contextual word representations [EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1909. 04164. pdf.
[49] LIU H, SINGH P. ConceptNet—a practical commonsense reasoning tool-kit[J]. BT Technology Journal, 2004, 22(4): 211-226.
[50] WANG X, GAO T, ZHU Z, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation [EB/OL]. [2021-01-05]. https://bakser. github. io/files/TACL-KEPLER/KEPLER. pdf.
[51] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph [C]// Proceedings of AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020: 2901-2908.
[52] WANG R, TANG D, DUAN N, et al. K-adapter: infusing knowledge into pre-trained models with adapters[EB/OL]. [2021-01-05]. https://arxiv. org/pdf/2002. 01808v3. pdf.
[53] ZHOU H, YOUNG T, HUANG M, et al. Commonsense knowledge aware conversation generation with graph attention[EB/OL]. [2021-01-05]. http://coai. cs. tsinghua. edu. cn/hml/media/files/2018_commonsense_ZhouHao_3_TYVQ7Iq. pdf.
[54] WU S, LI Y, ZHANG D, et al. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. San Diego, USA: Association for Computational Linguistics, 2020: 5811-5820.
[55] ZHOU H, YOUNG T, HUANG M, et al. Commonsense knowledge aware conversation generation with graph attention[EB/OL]. [2021-01-05]. http://coai. cs. tsinghua. edu. cn/hml/media/files/2018_commonsense_ZhouHao_3_TYVQ7Iq. pdf.
[56] HAYKIN S, KOSKO B. Gradientbased learning applied to document recognition[EB/OL]. [2021-01-05]. https://axon. cs. byu. edu/~martinez/classes/678/Papers/Convolution_nets. pdf.
[57] WEI J, REN X, LI X, et al. NEZHA: neural contextualized representation for Chinese language understanding[EB/OL]. [2021-01-05]. https://lonepatient. top/2020/01/20/NEZHA.
[58] SHAOUL C, BAAYEN R H, WESTBURY C F. N-gram probability effects in a cloze task[J]. The Mental Lexicon, 2014, 9(3): 437-472.
[59] DIAO S, BAI J, SONG Y, et al. ZEN: pre-training Chinese text encoder enhanced by N-gram representations[EB/OL]. [2021-01-05]. https://aclanthology. org/2020. findings-emnlp. 425. pdf.
[60] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT [EB/OL]. [2021-01-05]. https://arxiv. org/pdf/1906. 08101v2. pdf.
[61] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
[62] LEE J S, HSIANG J. Patent classification by fine-tuning BERT language model[EB/OL]. [2021-01-05]. https://arxiv. org/ftp/arxiv/papers/1906/1906. 02124. pdf.
[63] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[EB/OL]. [2021-01-05]. https://aclanthology. org/D19-1371. pdf.
[64] ZHAO S, GUPTA R, SONG Y, et al. Extreme language model compression with optimal subwords and shared projections[EB/OL]. [2021-01-05]. https://openreview. net/pdf? id=S1x6ueSKPr.

编辑 吴云芳