文章编号：1003-0077(2021)07-0010-20

自然语言预训练模型知识增强方法综述

孙 毅，裘杭萍，郑 雨，张超然，郝 超

（陆军工程大学 指挥控制工程学院，江苏 南京 210001）

摘 要：将知识引入到依靠数据驱动的人工智能模型中是实现人机混合智能的一种重要途径。当前以 BERT 为代表的预训练模型在自然语言处理领域取得了显著的成功，但是由于预训练模型大多是在大规模非结构化的语料数据上训练出来的，因此可以通过引入外部知识在一定程度上弥补其在确定性和可解释性上的缺陷。该文针对预训练词嵌入和预训练上下文编码器两个预训练模型的发展阶段，分析了它们的特点和缺陷，阐述了知识增强的相关概念，提出了预训练词嵌入知识增强的分类方法，将其分为四类：词嵌入改造、层次化编解码过程、优化注意力和引入知识记忆。将预训练上下文编码器的知识增强方法分为任务特定和任务通用两大类，并根据引入知识的显隐性对其中任务通用的知识增强方法进行了进一步的细分。该文通过分析预训练模型知识增强方法的类型和特点，为实现人机混合的人工智能提供了模式和算法上的参考依据。

关键词：预训练语言模型；知识增强；预训练词嵌入；预训练上下文编码器

中图分类号：TP391 **文献标识码**：A

Knowledge Enhancement for Pre-trained Language Models: A Survey

SUN Yi, QIU Hangping, ZHENG Yu, ZHANG Chaoran, HAO Chao

(Command and Control Engineering College, Army Engineering University of PLA, Nanjing, Jiangsu 210001, China)

Abstract: Introducing knowledge into data-driven artificial intelligence models is an important way to realize human-machine hybrid intelligence. The current pre-trained language models represented by BERT have achieved remarkable success in the field of natural language processing. However, the pre-trained language models are trained on large scale unstructured corpus data, and it is necessary to introduce external knowledge to alleviate its defects in determinacy and interpretability to some extent. In this paper, the characteristics and limitations of two kinds of pre-trained language models, pre-trained word embeddings and pre-trained context encoders, are analyzed. The related concepts of knowledge enhancement are explained. Four types of knowledge enhancement methods of pre-trained word embeddings are summarized and analyzed, which are pre-trained word embeddings retrofitting, hierarchizing the process of encoding and decoding, attention mechanism optimization and knowledge memory introduction. The knowledge enhancement methods of pre-training context encoders are described from two perspectives: 1) task-specific and task-agnostic; 2) explicit knowledge and implicit knowledge. Through the summary and analysis of the knowledge enhancement methods of the pre-trained language model, the basic pattern and algorithm are provided for the human-machine hybrid artificial intelligence.

Keywords: pre-trained language model; knowledge enhancement; pre-trained word embedding; pre-trained contextual encoder

0 引言

数据驱动和知识驱动是当前实现人工智能的两条重要途径。近年来以深度学习为代表的数据驱动方法在各类任务上取得了极大的成功，在自然语言处理领域，在大规模语料上训练的预训练语言模型在各项任务上均显示出良好的性能。尽管如此，当

收稿日期：2020-10-26 定稿日期：2021-01-17

基金项目：国防科技创新特区计划项目(1916311LZ001003)；装备发展部基金项目(6141B08010101)

前以数据驱动的方法仍然存在较大的局限性，张钹院士认为其适用场景仅限于：具有充分知识或数据、稳定性、完全信息、静态、特定领域与单任务。知识驱动的人工智能具有良好的逻辑性和可解释性，但却严重依赖人工定义的知识与规则，缺少对特征抽象和学习的能力，难以完整地表示人类的经验和知识。

通过将知识融入深度学习模型中，可以在一定程度上提高模型的泛化能力，同时增强模型的可控性。当前，从大规模数据中进行知识抽取和知识图谱建设的相关工作已逐渐成熟，但由于知识与训练语料或数据的异构性，对于知识赋能深度学习模型、指导应用实践的方法研究仍不够充分。因此，本文通过总结分析自然语言预训练模型知识增强的方法，来展示知识增强深度学习模型的途径和发展趋势，为实现通用的人机混合智能提供可借鉴的思路。

本文主要包括以下四个部分：

1）介绍了预训练模型与知识增强两个基本概念，分析了预训练词嵌入和预训练上下文编码器的特点与不足，阐述了知识的分类、知识库的类型和知识增强的基本模式。

2）总结分析了以预训练词嵌入为基础的语言模型知识增强的方法，提出将预训练词嵌入知识增强分为四类：词嵌入改造、层次化编解码过程、优化注意力和引入知识记忆，并分析了他们的特点与适用场景。

3）从任务特定和任务通用、引入知识的显隐性两个视角对预训练上下文编码器的知识增强进行了总结分类，针对通用预训练上下文编码器，着重从显性知识和隐性知识的角度对其知识增强方法进行了阐述。

4）对当前语言模型知识增强方法进行了进一步的总结分析，并为其模式和方法的发展方向进行了展望。

本文的文献主要来源于近年来自然语言处理与人工智能领域的顶级会议和刊物，如 ACL、AAAI、ICLR、IJCAI、EMNLP、NAACL、TACL 等。

1　基本概念及问题分析

本节主要对预训练模型和知识增强两个主要概念进行介绍，分析预训练词嵌入和预训练上下文编码器的特点和不足，并对知识增强中知识的分类、知识库的类型和知识增强的模式进行了阐述。

1.1　预训练模型

自然语言预训练模型（pre-trained language model，PLM 或 PTM）也被称为预训练语言模型或预训练模型，Qiu 等人[1]将预训练模型的发展分为预训练词嵌入（pre-trained word embeddings，PWE）和预训练上下文编码器（pre-trained contextual encoders，PCE）两个阶段，本文同样按照这样的划分方法。由于目前针对以上两个阶段的研究都有了较为完整和全面的综述[1-4]，且本文的着眼点为预训练模型的知识增强方法，因此本文不再对其中与本文侧重点相关度不大的知识点展开介绍。

1.1.1　预训练词嵌入及其存在问题分析

预训练词嵌入（也称词向量）技术是利用大规模的文本语料，将单词的语义嵌入到低维、稠密、长度固定的数值向量中的方法[3]，其构建过程主要遵循单词的分布假设和词的共现统计[2]。在词嵌入技术得到广泛应用之前，文本的特征向量通常是随机初始化的，即假设有“足够”的训练数据，可以把单词的特征向量认为是模型的一种参数，并且可以通过神经网络的训练将其调整为合适的向量表示[5]。对于训练数据不足的任务，可以把词嵌入技术看作是一种有效的引入外部知识的方法，知识的来源即大规模的语料库。

早在 2003 年，Bengio 等人就提出了 NNLM 模型[6]，但受限于模型的复杂度和当时的算力水平，NNLM 没有得到广泛的应用。2013 年 Mikolov 等人[7]通过优化 NNLM 模型提出 skip-gram 和 CBOW 模型，也就是 Word2Vec 模型，Word2Vec 不仅可以快速地在大规模语料上进行训练，并且可以很好地表征单词的语义[4]。此后，基于全局语料和词共现矩阵的 GloVe[8]、增加了单词 n-gram 信息的 FastText[9]等模型也相继出现。

尽管预训练词嵌入取得了较大的成功，但仍然存在以下不足：

（1）由于静态词嵌入中同一个词只对应一个嵌入表示，不能根据上下文调整词的表示，难以解决一词多义问题。

（2）浅层神经网络不足以捕捉到大规模语料中的复杂信息，且由于预训练的网络结构不复用，无法将词之间的依赖关系传递到下游任务中。

（3）部分罕见词和未登录词得不到充分的训练，对这些词的表示上存在误差。

其中，前两点由预训练词嵌入的网络结构的缺

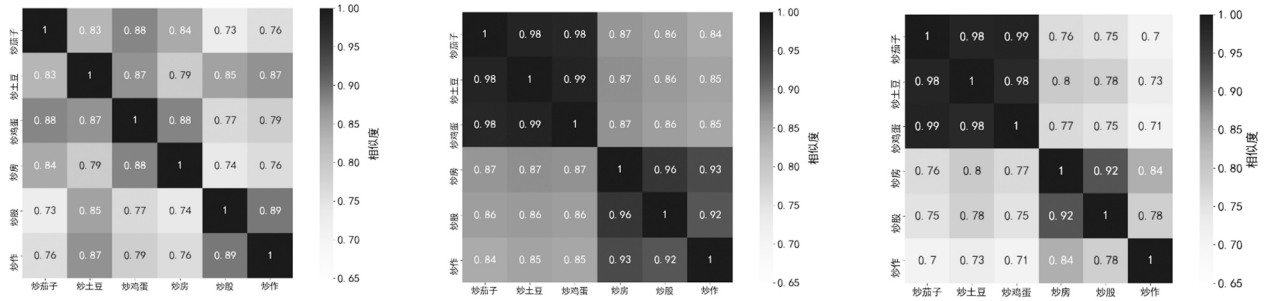
陷导致;第三点主要由训练语料决定,但可以通过模型、算法进行弥补。

1.1.2 预训练上下文编码器及其存在问题分析

CoVe[10]和ELMo[11]等上下文相关动态词嵌入模型改善了传统静态词嵌入在一词多义方面的缺陷。而BERT[12]等预训练上下文编码器最大的特点是采用了预训练-微调的方式,使用基于自注意力的全连接Transformer网络[13],增加了网络的深度,不仅使得模型能够更好地解决长距离依赖问题,同时由于网络足够深,编码能力足够强,文本的表示具有更高的动态性,可以更好地理解语义,进一步解决了一词多义问题[14-17]。

(1) 预训练上下文编码器与一词多义

例如,"炒"一词在百度百科中有两个基本释义:①一种烹饪方法,把食物放在锅里加热并不断翻动使熟;②为了扩大影响或制造轰动效应而反复地进行报道和宣传,利用价格的升降不断地买进卖出,以从中获利。如图1所示,从左到右依次为ELMo①(LSTM第1层与第2层取平均)、BERT②(倒数第2层)和RoBERTa[18](倒数第2层)对"炒"在6种不同的上下文中("炒茄子"、"炒土豆"、"炒鸡蛋"、"炒房"、"炒股"和"炒作")的向量表示间的余弦相似度,前3种和后3种不同的上下文分别对应炒的两个基本释义。不难看出,3个模型都能一定程度上区分这两种不同的含义,RoBERTa区分能力最强。



(a) ELMo (b) BERT (b) RoBERTa

图1 ELMo、BERT和RoBERTa动态性对比示例

(2) 预训练上下文编码器知识存储

当模型无法从上下文中判断文本的含义时,需要进一步引入其中所涉及的实体的相关属性、状态等背景知识,或是对该文本进行领域限定。当前的研究对预训练上下文编码器的知识存储能力尚存在争议,主要有以下两种观点:

观点1 预训练上下文编码器具备一定的存储能力

Petroni等人[19]通过完形填空的方式来检验预训练模型在知识存储方面的能力(language model analysis probe,LAMA probe)。该检验方法选取知识图谱(包括Google-RE③、T-Rex④和ConceptNet等)中的三元组信息,例如(Dante, born-in, Florence),将其转化为完形填空的形式"Dante was born in _____",让预训练模型进行预测,实验表明,当BERT给出候选的10个答案时,命中率接近60%,前100个候选答案的命中率接近80%,当仅给出1个答案时效果并不理想,但这已经可以表明预训练模型可以很好地存储一些较为通用的知识。同时,BERT在一对一关系预测效果上表现较好,而在多对多关系上表现较差。Roberts等人[20]通过微调预训练模型T5,在不检索任何外部上下文或知识的情况下来完成开放领域问答任务(Natural Questions[21]、WebQuestions[22]和TriviaQA[23]),实验表明,随着训练规模的提升,T5模型的效果也随之提升,甚至与当前最先进的检索式模型不相上下。

观点2 预训练上下文编码器知识存储有局限性

Kassner和Schütze[24]设计了negated LAMA,将LAMA[19]中的问题改为否定形式,例如"Dante was not born in _____",实验表明,未经微调的预训练模型难以应对否定形式的事实推理,甚至会得出"Birds cannot fly"的结论。Poerner等人[25]设计了LAMA-UHN(Un Helpful Names),对Google-RE和T-Rex进一步做了筛选:①删除三元组中客体是主体子串的情况(例如,Apple和Apple-Watch);②对人名进行拆分,删除姓或名有明显地

① https://github.com/HIT-SCIR/ELMoForManyLangs

② https://github.com/ymcui/Chinese-BERT-wwm

③ https://code.google.com/archive/p/relation-extraction-corpus

④ https://hadyelsahar.github.io/t-rex

域倾向的三元组。经过以上方法筛除较为简单的数据，BERT 模型的效果大幅度下降。Niven 和 Kao[26]认为 ACRT 数据集[27]存在词分布不均匀情况，给 BERT 带来了统计线索，例如，BERT 可以仅利用部分文本中的单词“not”就能推断出答案，在重新对数据集进行处理后，BERT 模型结果接近随机猜测。

预训练模型在不同的阶段获取了不同的知识，Kovaleva 等人[28]通过实验发现，即使仅使用 BERT 的结构，不进行预训练，使用随机初始的参数，也能在一些任务上取得较好的结果。根据上述研究本文认为，通常的预训练模型中所存储的知识可分为“模糊知识”和“精确知识”两类：

• “模糊知识”通常在预训练阶段获得，通过在大规模通用语料上进行训练，预训练模型通过文本的共现规律对大量常见的单词组合、通用实体和基本常识有较好的记忆，但对具体的实体及其属性很难形成精确的记忆，因此在预训练的过程中形成的知识记忆是相对“模糊”的。

• “精确知识”通常在微调阶段获得，模型通过微调捕捉到了具体任务中目标函数相关的“精确”的线索知识。

(3) 预训练上下文编码器与中文房间问题

Searle 于 1980 年提出了著名的中文房间问题[29]，假如将中文房间中的人与预训练模型做类比，预训练模型更像是一个有着一定计算能力的人，将所有手册进行反复地学习，钻研其中的统计规律并保存在模型参数中，当需要回答问题时便利用这些规律进行回答。当前预训练模型甚至还远远达不到中文房间以假乱真的效果，主要是尚不具备逻辑推理能力，例如，“否定推断”能力、“数字推导”能力、和“比较”能力等[30]。

1.2　预训练模型

预训练词嵌入模型和预训练上下文编码器通过在大规模语料上进行学习，在各项自然语言处理任务上取得了较大的成功。但受限于语料的规模、长尾现象及模型的学习能力等因素，当前的语言模型仍存在知识缺乏的问题。本文所研究的知识增强是指通过引入人工的知识信息，改善模型的不足、提升模型性能的相关方法。

1.2.1　知识

对知识的定义和分类，著名哲学家迈克尔·波兰尼在从哲学领域提出将知识分为显性知识(也称名言知识)和隐性知识(也称默会知识、非名言知识)。显性知识是能够用各种名言符号加以表述的知识，隐性知识是指知道但难以言传的知识。波兰尼指出：“在语言拓展人类的智力，使之大大地超越纯粹隐性领域的同时，语言的逻辑本身——语言的运用方式——仍然是隐性的。”[31-33]

本文所认为的知识信息包括显性知识和隐性知识两类：显性知识包括语义词典、语义网络和知识图谱等知识库中的词义解释、语义关系等明确的知识，这些知识通常被称为常识(commonsense)、世界知识(world knowledge)等；隐性知识指在模型训练过程中使用的掩码规则、分词、词性、语义角色、情感等难以明确但对模型理解文本有益的知识，这些知识一般是实践规律，需要在具体的场景产生和发挥作用。

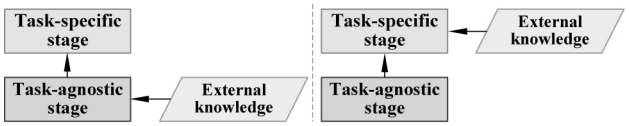
1.2.2　知识库

知识库(Knowledge Base 或 Knowledge Vault)是一个具有正式语义的数据集，它可以包含不同种类的知识，例如规则、事实、公理、定义、语句和基元等[34]。虽然知识库、本体(ontology)、知识图谱(knowledge graph)在定义和特点上有所不同[35]，但由于不是本文的重点，因此不过多对其区别进行讨论。本文将知识库的范围规定为以自然语言为基本形式由人类产生的具有一定规律的数据库或数据集。当前存在的各类知识库大多各具特色，其中主要包含两种类型的知识，一种是解释型的信息，例如，WordNet[36]中的词语解释和例句，BabelNet[37]中的百科词条，HowNet 中的概念(sense)和义原(sememe)；第二种是关系信息，最典型的是 ConceptNet①、Freebase②、Wikidata③ 等知识图谱中的知识三元组。

1.2.3　知识增强的模式

当前的语言模型大多分为两个阶段，一是与任务无关(task-agnostic stage)的预训练阶段，二是任务相关(task-specific stage)的阶段。因此对于语言模型知识增强的基本模式也可分为如图 2 所示的两类：一是在预训练阶段引入知识(a)；二是在任务相关阶段引入知识(b)。以上两种知识增强模式与人类的学习习惯类似。

① http://www.conceptnet.io
② freebase.com 已被关闭
③ https://www.wikidata.org



(a) 预训练阶段引入知识 (b) 任务相关阶段引入知识

图2 语言模型知识增强的两种基本模式

在任务无关的预训练阶段引入的知识通常覆盖范围较广，可以让模型较好地获得某个或某几个知识库中的全局知识。而在任务相关的阶段引入的知识，通常与具体的文本和任务密切相关，涉及到知识的检索，因此引入的知识更加精确，但也容易受限于局部的知识。

2 预训练词嵌入知识增强方法

预训练词嵌入知识增强的方法从不同的视角可以有不同的分类，例如可以按照知识库类型（语义词典、知识图谱等）进行分类。为了更好地总结其规律，本文从知识引入途径的角度将其分为四类：词嵌入改造、层次化编解码过程、优化注意力和引入知识记忆。此外，词嵌入模型在应用于下游任务时往往会结合不同的网络结构，因此对词嵌入的知识增强更准确的表达是对以词嵌入为基础的语言模型的知识增强。

2.1 词嵌入改造

预训练词嵌入改造（pre-trained word embeddings retrofitting）主要分为利用外部知识学习词义嵌入和优化词义嵌入分布两类。

2.1.1 学习词义嵌入

词义嵌入（sense embeddings）主要解决一词多义问题，其学习过程主要依靠知识库中的词义解释和同义关系等信息，词义嵌入与传统词嵌入的区别在于每个单词的词嵌入都有多个，在训练或使用过程中每个单词的表示被动态地选择为与上下文最接近的词义。

Chen等人[38]首先用Word2Vec学习单一词汇表示，然后将WordNet作为词义来源，用每个词义对应的解释中词汇嵌入的平均值作为词义的初始嵌入，通过修改Skip-gram的训练目标，学习WordNet中的词和多义词的嵌入，在使用时判断该词属于哪个含义。

Chen等人[39-40]提出CNN-VMSSG模型，该方法同样是首先利用Word2Vec从大规模语料中学习单词的嵌入表示，然后利用卷积神经网络将WordNet中词汇的解释内容初始化为单词的词义嵌入。最后，将这些初始化的词义嵌入输入Neelakantan等人[41]的MSSG（multiple-sense skip-gram，多义skip-gram）模型中以学习基于知识的词义嵌入。

Rothe和Schütze[42]提出的AutoExtend模型利用WordNet将标准的词嵌入扩展到其同义词集嵌入和语义嵌入。其训练过程中遵循两个目标准则：①词嵌入是其对应的所有词义嵌入之和；②同义词集嵌入是其中包含的所有词义嵌入之和。

Pilehvar等人[43]提出DeConf模型，将WordNet中的同义关系看作为一个图结构，其中的节点为WordNet同义词集中的一个词，边代表词之间的语义关系。用PageRank算法在WordNet的语义图中找出每个多义词对应的经过排序的语义偏置词集（sense biasing words），然后通过每个词义所对应的语义偏置词集来计算该语义的嵌入表示，如式（1）所示。

$$\vec{w}_{si}^{*}=\frac{\alpha\vec{w}_{si}+\sum_{bij\in B_i}\delta_{ij}\vec{w}_{bij}}{\alpha+\sum_{j}\delta_{ij}} \tag{1}$$

其中，$\vec{w}_{bij}$ 为词义对应的语义偏置词，$\delta_{ij}=e^{-\lambda r(i,j)}/|B_i|$，$r(i,j)$ 为语义偏置词的排序，B_i 为语义偏置词集合，α 与 λ 均为超参数。

2.1.2 优化词嵌入分布

词嵌入的分布情况决定了词嵌入的质量，受限于训练语料或是训练模型结构（窗口大小、网络深度等）等因素，词嵌入的分布情况往往不能满足不同任务的需求，因此部分研究通过外部知识对词嵌入的分布进行优化。

Faruqui等人[44]提出的基于语义关系的词嵌入改造，根据新的目标函数调整现有词嵌入的值，是词嵌入的一种后处理方法（post-processing pre-training）。其思路是对于给定的词嵌入，使具有同义关系的词尽量靠近，优化目标为最小化式（2）。

$$\sum_{i=1}^{|V|}(\alpha_i\|\vec{w}_i-\vec{w}_i^{*}\|+\sum_{(\vec{w}_i,\vec{w}_j)\in N}\beta_{i,j}\|\vec{w}_i-\vec{w}_j\|) \tag{2}$$

其中，$|V|$ 表示词汇表的大小，N 表示单词对形式的语义网络，$\vec{w}_i$ 和 $\vec{w}_j$ 表示同义词对嵌入，$\vec{w}_i^{*}$ 表示调整后的词嵌入，α_i 和 $\beta_{i,j}$ 为两组超参数。

Mrkšić等人[45]提出Attract-Repel模型，该模型通过BableNet中的大规模的多语言同义、反义关系，将高资源语言的语义转换到低资源语言，以改进资源较少的语言的单词表示效果。

Speer 等人[46]利用 ConceptNet 5.5 中的概念间的关系通过 PPMI 算法[47]得到 ConceptNet-PPMI 词嵌入，然后利用 ConcepNet 中的关系对 Word2Vec 和 GloVe 进行词嵌入改造，并通过 Speer 和 Chin[48]的扩充改造方法(expanded retrofitting)，最终得到词嵌入 ConceptNetNumberbatch，该词嵌入可以更好地反映词间关系，其词汇表也大大丰富。

2.2　层次化编解码过程

在 HowNet 中，义原被认为是汉语中最基本的、不易于分隔的最小语义单位。层次化编解码过程大多是利用了词—词义—义原间的层次化关系，如式(3)所示。在编解码过程中，不再仅对词进行操作，还会对词对应的词义和义原进行编码或预测，以达到将词义和义原间的关系知识引入到模型中的效果。

$$w_i \propto \sum_{\text{sense}_j \in w_i} \text{sense}_j \propto \sum \sum_{\text{sememe}_k \in \text{sense}_j} \text{sememe}_k \tag{3}$$

孙茂松等人[49]提出了融合 HowNet 和大规模语料库的义原向量学习神经网络模型。该模型首先使用 CBOW 模型在训练语料库上学习词的嵌入表示，然后固定训练好的词嵌入不变，训练词所辖的义原向量去逼近该词的嵌入，最后使得学习到的义原向量可以较好地预测所定义的词嵌入。实验结果表明，该模型能有效提升在词相似度和词义消歧任务上的性能，有助于低频词和多义词的处理。

Mancini 等人[50]提出了 SW2V 模型(senses and words to vectors)，先定义了语义网络，将同义词集视为节点，词之间的同义关系视为边，通过词-词义连接算法将每个词与其相关的词义集联系起来；然后利用 CBOW 模型，将输入改变为单词的上下文和上下文每个单词对应的所有词义，预测目标是这个被预测的单词和该单词对应的词义。

Niu 等人[51]基于 Word2Vec 中的 Skip-Gram 模型，提出了 SAT(sememe attention over target model)模型。与 Skip-Gram 模型只考虑上下文信息相比，SAT 模型同时考虑单词的义原信息，使用义原信息辅助模型更好地理解单词词义。具体做法是：根据上下文单词对中心词进行词义消歧，使用注意力机制计算上下文对该单词各个词义的权重，然后使用词义嵌入(sense embedding)的加权平均值表示单词嵌入。

为优化词典扩展任务，Zeng 等人[52]利用大规模文本数据学习每个词语的分布式嵌入表示，然后用 LIWC(linguistic inquiry and word count)词典单词作为训练数据训练分类器，并用 HowNet 提供的义原标注信息构建义原注意力以提升 Seq2Seq 模型对单词的层次分类效果。该模型将单词所对应的义原作为上下文输入到 LSTM 模型的编码层中。

传统语言模型在编码输入序列后直接在词层面或字层面进行预测，Gu 等人[53]通过引入 HowNet 中“词-词义-义原”的结构关系，层次化 Seq2Seq 的预测过程，提出 SDLM(义原驱动的语言模型)，进而提高语言模型的性能和可解释性。义原驱动的解码器以循环神经网络输出的上下文向量作为输入，输出预测下一个单词的概率，其结构包括以下三个层次化的模块：①给定循环神经网络最后生成的上下文向量，预测每个义原将在下个词中出现的概率；②使用上下文向量和第一层中的预测，给出每个词义出现的概率；③将第二层中的词义出现的概率边缘化得到每个单词的概率。

2.3　优化注意力

通过外部知识优化模型注意力，主要是利用外部知识来让模型能更好地注意到任务相关的文本内容。

Chen 等人[54]利用 WordNet 中词之间的同义、反义和上下位关系来辅助判断注意力机制中的权重，通过影响权重使神经网络对不同的词施加不同的注意力，以此来达到引入外部知识的目的。

Zou 等人[55]认为基于注意力的情感分析模型大多没有充分利用情感词汇，这些情感词汇提供了丰富的情感信息，在情感分析中起着至关重要的作用。通过情感词典指导注意力机制，使得情感分类更加关注情感词而不是特定领域的名词，将损失函数与句子中的情感词挂钩，让模型学习与情感词相关的注意力，使得模型能够更好地泛化到不同的领域中。

2.4　引入知识记忆

为模型引入外部知识记忆是几种方法中最为主流的知识增强方法。其核心方法是将外部知识库中的常识信息等信息编码进模型中，其本质是为模型引入了新的上下文信息，类似于开放领域问答(open question answer，OpenQA)，但由于知识的引入往往缺乏监督数据因而有所不同。如图 3 所示，引入知识记忆的步骤主要为：检索、编码、过滤和融合。

• 检索　由于外部知识库通常规模庞大，因此大多数模型通过设置规则并根据上下文及当前词对所需的知识进行检索，得到一定数量的候选知识。

• 编码　外部知识通常利用文本进行描述，因此需要编码器将其编码为知识记忆向量，该编码器可以是与任务编码器不同的，也可以是与任务编码器相同的，即对称孪生结构。

• 过滤　检索步骤中的规则设置相对简单，得到的候选知识数量较多且存在一定知识“噪声”，因此需要对候选知识重新赋予权重。该过程通常根据上下文或某个时刻的状态利用注意力机制进行计算，主要有如图4所示的三种类型：有监督、无监督和弱监督，其中有监督知识过滤通常需要对知识的过滤进行数据标注，无监督知识过滤则是利用任务文本与知识文本间的语义相似度等信息，弱监督知识过滤需要通过任务相关的监督信息来间接地对知识过滤过程进行优化。

• 融合　由于知识的编码方法与任务文本的编码方式可能是不同的，因此需要将知识向量融合到任务网络中。同时还需要考虑外部知识向量与原有向量融合时各自所占比重，通常通过设置超参数或利用门函数来实现。

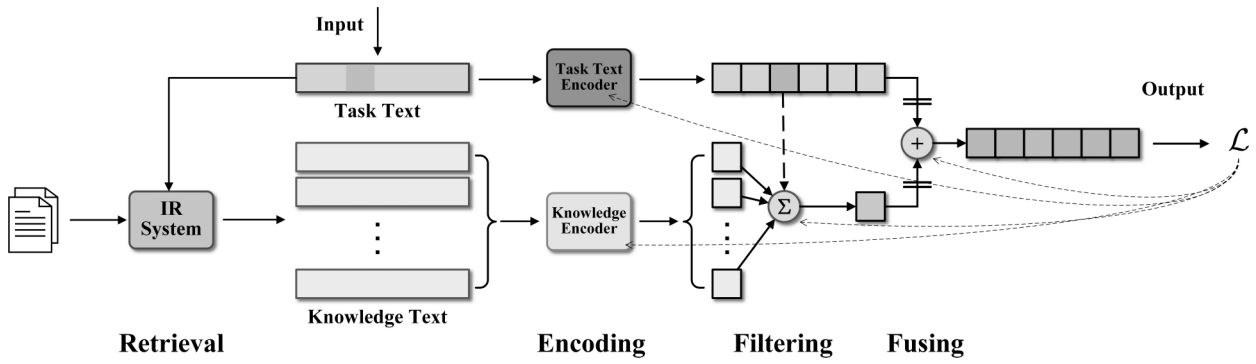
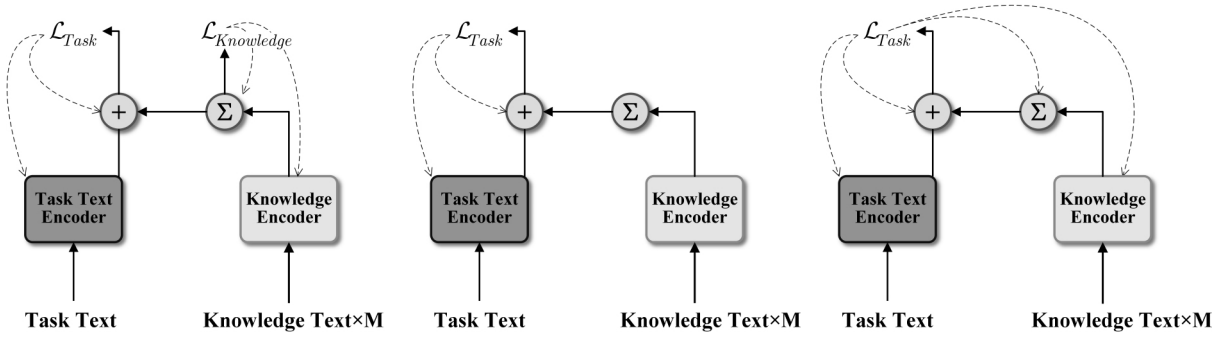


图3　知识记忆增强的语言模型的一般过程



(a) 有监督知识过滤　(b) 无监督知识过滤　(c) 弱监督知识过滤

图4　知识记忆增强中知识过滤的三种监督类型

Ghazvininejad等人[56]为解决在对话系统中模型缺乏真实的信息和背景实体的问题，提出利用维基百科、IMDB等知识库为对话引入其中的实体描述，即外部事实(external facts或world facts)。该模型(MTask)针对每段对话及其历史，通过基于关键词的检索引擎检索相关的外部知识，然后将外部事实与对话历史同时编码到循环神经网络中作为输入特征。

Parthasarathi和Pineau[57]利用Wikipedia summary①和NELL KB[58]的外部知识来提升对话性能，与Ghazvininejad等人不同的是，其模型Ext-ED(Extended Encoder-Decoder)在编码外部知识时使用了GloVe词嵌入的叠加均值来生成固定长度的外部上下文向量。

相比于以往的研究，为了更好地实现外部知识的引用，Facebook的Dinan等人[59]带来了更加工业化的解决方案。数据方面，他们首先利用“维基百科巫师”(Wizard Of Wikipedia)系统来创建基于外部知识检索的对话数据集。模型方面，在知识检索的步骤使用了Chen等人[60]的对文档基于TF-IDF权

① https://thijs.ai/Wikipedia-Summary-Dataset/

重的 bag-of-word 和 n-gram 向量的哈希检索方法，将知识库中检索到的与首轮对话和最近两轮对话最接近的文章作为外部知识；之后通过注意力机制对经过 Transformer 编码后的外部知识进行有监督选择，并将选择后的外部知识和对话上下文输入到端到端的 Transformer 模型中得到对话的回复。其中的 Transformer 是经过 Reddit 对话数据[61]预训练的。

Meng 等人[62]提出参考感知网络（reference-aware network，RefNet），该模型同时结合生成式方法和抽取式方法克服单个使用其中一个方法在对话系统中带来的缺陷。该模型中包含四个模块，分别是知识编码器、语境编码器、解码选择器和混合编码器。在解码的每个时间步，解码选择器都会在参考解码（referencing decoding）与生成解码（generationdecoding）之间做出选择，根据解码选择器的决定，混合解码器从背景知识中抽取一个完整语义单元或生成一个词，最终生成高信息量同时又不失流利的回复。

Yang 等人提出 KBLSTM 模型[63]，该模型根据上下文通过 BILINEAR 模型[64]来计算实体所对应的每个含义（即实体对应的三元组中客体）的权重，起到实体消歧的作用，然后将其编码到 LSTM 结构中来提高阅读理解的效果。

Young 等人[65]提出了一种用常识增强的端到端对话系统，根据对话选择正确的回答，在引入外部知识时首先根据对话检索相应的知识三元组，即断言（assert），然后通过 Tri-LSTM 对三元组按照从左到右的顺序依次编码主体、关系和客体，再利用训练集中的标准回复训练其与编码后的知识表示的匹配分数（match score），最后根据上下文及知识与候选回复的匹配程度选择最佳答案。

Mihaylov 和 Frank[66]提出了一种基于外部常识知识增强的完形填空式阅读理解模型（cloze-style reading comprehension）。该模型较 Young 等人[65]的工作更加细致，首先根据问题、上下文和候选答案对 OMCS 中的事实三元组进行知识检索，并根据匹配程度进行排序，然后在对知识进行编码时，按照从左到右的方式，使用 GRU 依次编码三元组中的主体、关系和客体，Mihaylov 和 Frank 认为这样做的目的有三个：①将知识与普通文本编码到同样的向量空间；②保留三元组的方向信息；③使用三元组中的关系去过滤主体中的信息来初始化客体。最后通过键值检索（key-value retrieval）[67]的方法，将三元组信息通过注意力机制叠加为一个单独知识记忆（knowledge memory）向量表示。基于以上的外部知识，阅读理解性能得到大幅提升。

Zhong 等人[68]为了优化对话中回复的情感识别问题，提出 KET（knowledge-enriched transformer）框架，对于对话中的非停用词，从 ConceptNet 中抽取相应的概念及其邻居节点组成的知识图。对于每个概念，从情感词典 NRC_VAD 中抽取情感值，之后利用动态上下文感知情感图注意力来计算融合了知识的上下文表示，然后通过层次自注意力机制来表示上下文，最后通过上下文与回复的交叉注意力来预测回复的情感。

Zhou 等人[69]提出了常识知识感知对话模型（commonsense knowledge aware conversational model，CCM），将 ConceptNet 中的知识三元组看作知识图，基于实体与相邻实体间的关系，获得结构更为清晰、语义更加连贯的知识图编码信息。该模型中设计了两种新的图注意力机制：①静态图注意力机制，对检索到的知识图进行编码，来提升问题的语义，帮助更充分地理解问题；②动态图注意机制，读取每个知识图及其中的三元组，然后利用图和三元组的语义信息来生成更合理的回复。

李强等人[70]将翻译模板内嵌到端到端的神经机器翻译模型中，提出了模板驱动的神经机器翻译模型，该模型通过使用额外的模板编码器对翻译模板进行端到端建模。使用知识门阀和注意力门阀，动态地控制解码过程中不同来源的知识对当前解码词汇的贡献度的大小。知识门阀的主要作用是对源语言句子和翻译模板的信息进行有效的表示，从而更好地对解码器进行初始化。注意力门阀的作用是动态地控制当前翻译词汇接收源语言句子或者翻译模板信息的多少。

3　预训练上下文编码器模型知识增强

预训练—微调是预训练上下文编码器给语言模型带来的新的范式，因此对于预训练上下文编码器的知识增强根据知识引入的阶段可分为：在预训练阶段引入知识、在微调和推理阶段引入知识。本文假定模型在微调和推理阶段是同构的，因此知识的引入在微调和推理阶段是同步的。

图 5 为本文总结的预训练上下文编码器模型知识增强的分类方法。首先，根据知识增强后模型的通用性可分为预训练上下文编码器特定任务知识增

强和通用预训练模型知识增强；在任务通用的模型中，可以根据融合知识的显性和隐性进行分类；在融合显性知识的模型中可根据引入知识的阶段（预训练阶段、微调推理阶段）和结构（同构知识、异构知识和关系知识）进行分类；由于大部分融合隐性知识的模型都是在预训练阶段进行的，因此根据引入知识的类型（掩码策略、实体替换、实体感知、语义角色和情感信息）进行分类。

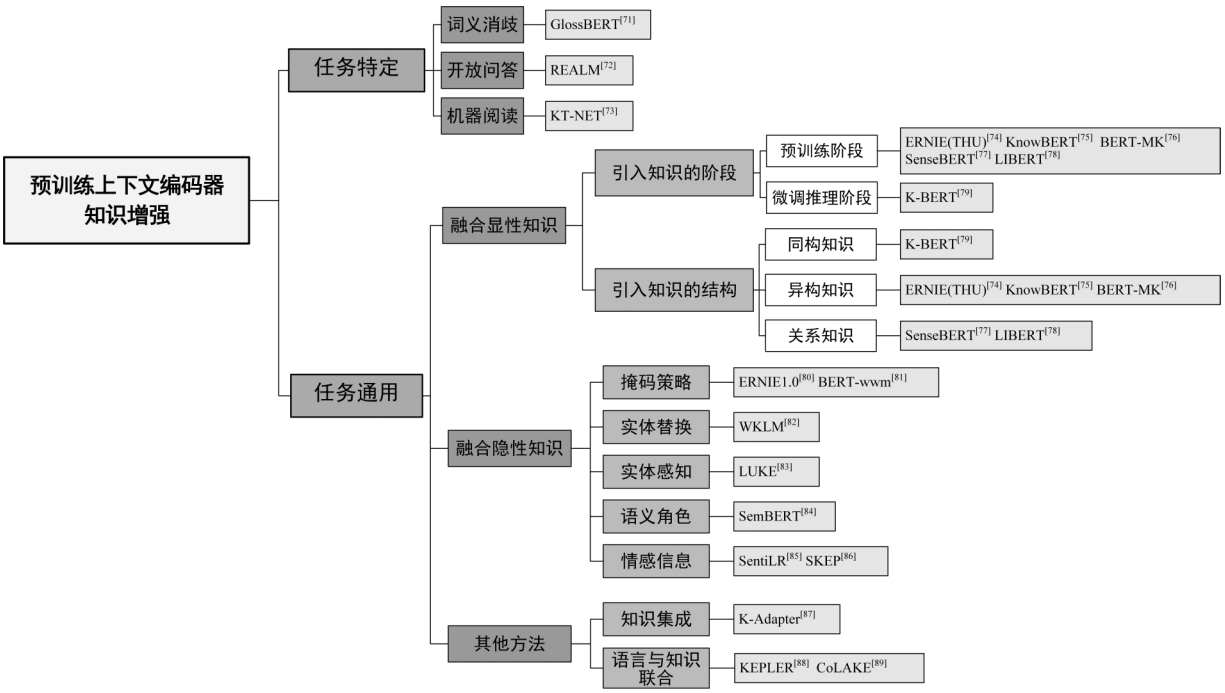


图5 预训练上下文编码器知识增强分类图

表1为预训练上下文编码器知识增强模型在模型初始化方法、预训练任务、知识形式、引入知识阶段、训练语料、知识源和所针对的特定任务或领域等方面的详细比对。

表1 预训练上下文编码器知识增强模型

模型名称	初始模型	预训练任务	知识形式	引入知识阶段	语料*	知识源	特定任务或领域
任务特定							
GlossBERT[71]	BERT	WSD	Word Gloss	Pre-Training	/	SemCor 3.0	WSD
REALM[72]	BERT	MLM	Textual Knowledge Corpus	Pre-Training & Fine-tuning &Inference	CC-News	English Wikipedia	Open QA
KT-NET[73]	BERT	MRC	Triples	Fine-tuning &Inference	/	WordNet+NELL	RC
任务无关 / 显性知识							
ERNIE (THU)[74]	BERT	MLM+NSP+dEA	Entity Embeddings	Pre-Training	English Wikipedia	Wikidata	/
KnowBERT[75]	BERT	MLM+NSP+EL	Entity Embeddings	Pre-Training	/	Wikipedia + CrossWikis +YAGO dictionary+ WordNet	/
BERT-MK[76]	/	MLM+KE	Knowledge Sub-Graph	Pre-Training	PubMed abstracts + PubMed Central full-text papers	UMLS	Medical#
SenseBERT[77]	BERT	MLM+SLM	Word Sense	Pre-Training	N/A	WordNet	/
LIBERT[78]	/	MLM+NSP+LRC	Lexical Relation	Pre-Training	English Wikipedia	WordNet+Roget's Thesaurus	/

续表

模型名称	初始模型	预训练任务	知识形式	引入知识阶段	语料*	知识源	特定任务或领域
K-BERT[79]	BERT	/	Triples	Fine-tuning&Inference	WikiZh+WebtextZh	CN-DBpedia + HowNe+MedicalKG	Finance/Law/Medicine#
任务无关 / 隐性知识							
ERNIE1.0[80]	/	MLM+NSP(DLM)	Phrase+Entity	Pre-Training	Chinese Wikepedia+Baidu Baike+Baidu news+Baidu Tieba	Tokenizer(Internal Tools)	/
BERT-wwm[81]	BERT	MLM+NSP	Word	Pre-Training	Wikipedia dump	LTP Tokenizer	/
WKLM[82]	/	MLM+WKLM	Entity	Pre-Training	BooksCorpus+English Wikipedia	English Wikipedia+Wikidata	/
SemBERT[83]	BERT	/	Semantic Role Label	Fine-tuning&Inference	/	Semantic Role Labeler	/
SentiLR[84]	BERT、RoBERTa	LA-MLM+NSP	Polarity+POS+Sentiment	Pre-Training	Yelp Dataset Challenge 2019	SentiWordNet	/
SKPE[85]	RoBERTa	SW+WP+AP	Sentiment	Pre-Training	N/A	Sentiment Seed Words	Sentiment Analysis#
LUKE[86]	RoBERTa	MLM+PME	Entity	Pre-Training	Wikipedia	Wikipedia	Entity-Related Task#
任务无关 / 其他							
K-Adapter[87]	RoBERTa	FA+LA	Factual Knowledge+Linguistic Knowledge	Pre-Training	/	T-REx-rc+Dependency Parser	/
KEPLER[88]	RoBERTa	MLM+NSP	Triples	Pre-Training	BookCorpus+English Wikipedia	Wikidata5M+WordNet	/
CoLAKE[89]	RoBERTa	MLM(MWD+MEN+MRN)	Knowledge Sub-Graph	Pre-Training	English Wikipedia	Wikidata5M	/

注："*"表示此处语料库为除去初始参数的模型所使用的语料库的部分；"/"表示无相关初始模型、语料或特定任务或领域；"N/A"表示原文中没有明确(not available)；"#"表示模型从结构上是通用的，但引入的知识主要针对某类任务。

在对预训练上下文编码器进行知识增强时，通常会引入新的与知识相关的预训练目标，因此有必要简单介绍预训练模型的基本任务。根据本文的研究侧重点，从词和句子两个粒度介绍预训练上下文编码器的基本任务：

• 在词的层面上，常见的训练任务有屏蔽语言模型(masked language modeling，MLM)和单词替换检测(replaced token detection，RTD)。MLM 模型即常说的掩码模型，是将部分单词用特殊符号[MASK]进行遮挡，令模型对这部分单词进行预测，BERT[12]和 RoBERTa[18]等模型都采用了该任务；RTD 将部分单词替换成其他单词，然后令模型预测单词是否被替换掉，ELECTRA[90]模型采用了该方法。

• 在句子的层面上，常见的预训练任务有下一句预测(next sentence prediction，NSP)和句子顺序预测(sentence order prediction，SOP)。NSP 将部分原始句子对中的后一个句子进行随机替换，令模型判断第二个句子是不是第一个句子的后续句子，BERT 模型采用了这个任务；SOP 使用同一文档中的两个连续片段作为正样本，而相同的两个连续片段互换顺序作为负样本，ALBERT[91]采用了这个任务。

3.1　预训练上下文编码器特定任务知识增强

针对特定任务而设计的知识增强方法可以有效地提高模型在特定任务上的性能，其特点是专用性强，但基本上不适用于其他类型的任务。

BERT 良好的结构为背景知识的引入提供了方便，当引入非结构化的文本知识时，可以灵活地应用 BERT 的编码层和一些特殊符号(如[SEP]等)。为提高 BERT 对词义消歧的效果，Huang 等人[71]提出了 GlossBERT，将 WordNet 的词语解释加入到 BERT 的输入中，对于每个需要消歧的词，分别输入 WordNet 中相关的 4 个词义解释，其中仅有一个是符合当前语境的，输出标签为单词是否与该解释相

匹配。此外，GlossBERT中还加入了一定的弱监督信号，使得模型能更好地锁定目标词。

在传统的语言模型中，引入背景知识通常是在具体的任务中，而对于预训练模型，在预训练阶段同样引入背景知识是一种比较自然的想法。Guu等人[72]认为预训练模型所学到的东西只是隐式的保存在模型的众多参数中，并不能以一种直观的方式表明模型是否真正地学到了，以及它学到了什么。因此，他们提出REALM模型，在预训练和微调两个阶段都增加了知识检索的步骤。预训练时，模型利用MLM预测MASK的区域正确的内容前，首先通过检索模型从文档集中检索相关的文档，然后利用检索结果中TOP-K文档中的内容来进行正确的预测。微调阶段，模型同样是先从完整的文档集中进行相应的检索，最后利用检索结果来完成开放领域问答任务，从而实现对于模型的微调。

Yang等人[73]提出的KT-NET，是利用BERT模型对KBLSTM[63]的升级，即将编码层替换为BERT，同样是引用BILINEAR[64]对检索到的三元组知识进行编码，将相关知识引入到阅读理解中。

3.2 通用预训练上下文编码器知识增强

通用预训练上下文编码器知识增强是将知识引入到预训练模型中，同时保持了模型的通用性；或者为模型引入某一类特定的知识，在不影响模型在通用任务上性能的前提下，提高部分任务的性能。受到波兰尼认知理论的启发，本文主要将其分为融合显性知识和融合隐性知识两大类。

3.2.1 融合显性知识的预训练上下文编码器

显性知识包括语义词典、语义网络和知识图谱等知识库中的词义解释、语义关系等明确的知识，这些知识通常被称为常识、世界知识等。预训练模型在大规模语料库上进行预训练时已经掌握了与此相似的知识，通过进一步从知识库中引入显性知识，可以强化模型中的知识信息和关系约束，使得知识更加精确。

从知识引入的阶段来看，融合显性知识的预训练上下文编码器可分为两个阶段：①预训练阶段，如ERNIE(THU)、KnowBERT等[74-79]，这些模型在预训练阶段将知识引入到模型中，通过各类与知识相关的目标函数强化对知识编码的能力；②微调和推理阶段，如K-BERT[79]仅在微调和推理阶段将知识引入，通过丰富上下文的语境，让模型更好地对实体进行编码。

在预训练阶段引入知识的模型中，由于模型结构的不同，初始化方法也各不相同。通常情况下，大部分模型利用已经过训练的预训练模型进行初始化，对模型的训练有良好的帮助。同时，为了防止模型因知识相关任务而带来的遗忘问题，仍需要预训练模型的基本任务(如MLM和NSP等)进行多任务训练。

根据引入知识的结构的不同，可分为三类：①引入同构知识，同构知识指引入与预训练模型原本训练数据形式相同的知识，即自然语言文本的形式，例如K-BERT[79]将知识图谱中的知识三元组以自然语言的方式插入到原文中；②引入异构知识，例如ERNIE(THU)[74]和KnowBERT[75]中的实体嵌入、BERT-MK[76]中的知识图表征；③引入关系约束，这类方法通过设计例如分类任务的方法，来实现引入某类关系知识，例如SenseBERT[77]中引入的一词多义知识(可以看作是词与超义的关系)和LIBERT[78]中的上下位关系。

1) ERNIE(THU)

清华大学Zhang等人的[74]ERNIE利用TransE[92]来编码知识图谱中的事实信息，设计了新的预训练目标dEA(denoising entity auto-encoder)，即随机掩盖输入中的部分命名实体，使模型从知识图谱中预测出合适的实体，并且在预测过程中引入背景知识信息。模型主要由两层组成，分别是T-Encoder(文本编码层)和K-Encoder(知识编码层)。其中T-Encoder主要是对输入的文本进行编码，提取词法及语义的信息。K-Encoder主要进行的是知识实体的表示以及知识融合。由于大部分数据集都没有相关的实体标注，对于文本中出现的实体，使用TAGME[93]将其链接到知识图谱中。

2) KnowBERT

KnowBERT[75]将BERT的输入文本中提及的实体通过指代消解和实体链接的方法链接到知识库中(维基百科和WordNet)。对于不包含图结构的维基百科，使用skip-gram[94]编码实体相关文章的标题描述(如对于歌手“Prince”，其对应的维基百科标题为“Prince (musician)”)；对于包含图结构的WordNet，通过知识图谱嵌入方法TuckER[95]编码其中的实体(包含其中关系和解释信息)。将输入根据知识信息重新上下文化，以使其携带实体信息。通过知识库的引入，解决因不频繁出现的常识或者

长距离依赖造成的难以学习选择偏好的问题。KnowBERT 存在的问题在于需要有监督的带有实体链接数据集，如 CrossWikis[96] 和 YAGO dictionary[97]。

3）BERT-MK

ERNIE(THU)中的实体嵌入通过 TransE 学习得来，但 TransE 只是将知识图谱中的每一个三元组看作一个训练实体，不足以对知识图中节点之间的复杂信息关系进行建模。在医学知识图谱中，有些实体有大量的相关邻域，而 TransE 不能为相应的邻域建模。因此 He 等人[76]提出一种能对任意子图进行建模的知识表征学习方法 KG-Transformer，该方法极大地丰富了知识表征中的信息量，探索了实体和关系的联合学习。将图的上下文知识整合到模型中来提高模型在医学领域的性能。

4）K-BERT

在公共领域上进行预训练的模型通常缺乏专业领域的知识。为使模型在处理领域文本时，可以用到相关的知识，Liu 等人[79]提出知识赋能的语言表示模型 K-BERT。该模型将知识三元组插入到句子中作为领域知识，同时为避免因外部知识的插入造成句子结构信息丢失的问题，该模型提出了软位置(soft-position)、可见矩阵(visible matrix)和遮蔽注意力(mask-Transformer)，使得原文与外部信息共同构成句子树，而树的枝干之间互不影响。K-BERT 在开放领域任务上有小幅的提升，在特定领域任务上性能提升明显。

K-BERT 有较高的易用性，但仍存在一些问题：①并没有对三元组的获取进行过滤；②由于预训练模型已经包含部分公共领域知识，这些知识的引入对模型理解文本的帮助并不大，甚至有可能会增加噪声信息。

5）SenseBERT

为使模型可以更好地学习词义，解决一词多义问题，Levine 等人[77]提出 SenseBERT，除了基础的 MLM 任务，将 WordNet 中的超义(supersense，如单词 chocolate 对应超义 noun.food，noun.attribute 等)信息引入到模型中，提出语义层面语言任务模型(semantic-level language model)，即令模型预测被掩盖的词所对应的超义。该模型在词义消歧任务上的性能取得了显著的提升，模型可以更好地预测词语在给定语境下的实际含义。但 SenseBERT 在通用任务上效果提升并不明显，甚至在个别任务上有一定的性能下降。

6）LIBERT

为增加预训练模型在语义相似性上的约束，Lauscher 等人[78]将词汇关系(同义关系和上下位关系)分类任务 LRC(lexical relation classification)作为新的预训练目标，与 BERT 原有的 MLM 和 NSP 任务交替训练模型。LRC 在形式上与 NSP 一致，且在内涵上有相似性，因此 LIBERT 模型在绝大部分通用任务上都取得了性能提升，说明语义相似性约束不仅对静态词嵌入生成有着积极的作用[98]，而且对于预训练模型有通用的价值。

3.2.2　融合隐性知识的预训练上下文编码器

隐性知识指在模型训练过程中使用的掩码规则、分词、词性、语义角色、情感等难以明确但对模型理解文本有益的知识。与显性知识的“授之以鱼”不同，隐性知识的引入更像是“授之以渔”，通过各种启发式的规则使得模型能够更加完整地学习语义单元与真实世界的实体关系。令语言模型能够更好地理解、学习语料的内在知识。

1）ERNIE 1.0

相比于 BERT，ERNIE 1.0[80]改进了两种掩码策略，一种是基于短语的，另外一种是基于实体的。在 ERNIE 中，将由多个汉字组成的短语或者实体当成一个统一单元，相比于 BERT 基于字的掩码方式，这个单元中的所有字在训练的时候会一起被掩盖。与直接将知识类实体映射成向量相比，ERNIE 通过统一掩码的方式可以学习到潜在的知识及更长的语义依赖，以此让模型更具泛化性。ERNIE 2.0[99]同样沿用这一知识掩码任务(knowledge masking task)。

2）BERT-wwm

全词掩码(whole word masking，WWM)的 BERT 是谷歌 BERT 的升级版本，升级内容主要是更改了原预训练阶段的训练样本生成策略。在谷歌官方发布的中文 BERT 模型中，中文以字为粒度进行切分，没有考虑到传统中文自然语言处理中的分词现象。Cui 等人[81]将全词掩码的方法应用了中文 BERT 中，对中文维基百科语料使用了哈工大分词工具 LTP①[100]进行分词，即对组成同一个词的汉字全部进行掩码。类似地，华为的 NEZHA 模型[101]通过 jieba 分词②来实现全词掩码以提升模型性能。

① http://ltp.ai

② https://github.com/fxsjy/jieba

3）WKLM

Xiong等人[82]的WKLM模型，提出弱监督训练方法ERT（entity replacement training），利用实体链接和维基百科，将文本中出现的实体随机替换为同类别的其他实体，让模型预测实体是否被替换，以此来提高模型从文本中捕捉真实世界实体知识的能力。

4）LUKE

为使预训练上下文编码器可以更好地感知文本中的实体，Yamada等人[83]提出LUKE模型。该模型在预训练的过程中，针对存在实体的文本增加独立的实体输入，并为之设置实体预测的任务（predicting the masked entities，PME），使模型能够更好地建模文本中同时出现的多个实体。同时，由于需要处理单词和实体两种不同的输入，模型引入了实体感知的自注意力机制（entity-aware self-attention mechanism）。通过在预训练阶段强化实体信息，模型在多个实体相关的测试任务上取得了较基线模型RoBERTa更好的效果。

5）SemBERT

为给语言模型添加结构化的语义信息，构建更加精确的文本表示，提高语言理解能力，Zhang等人[84]提出语义感知模型SemBERT。通过语义角色标注（semanticrole labling，SRL）工具标注输入文本的谓词、论元等语义角色，然后通过语义集成模块将BERT的文本表示与语义标签向量相融合，以得到可用于下游任务的联合表示。SemBERT保持了BERT的易用性，只需要进行适应性的微调，而无须对模型进行大幅度的修改，并在多数自然语言任务特别是语义推断任务上性能取得了明显提升。

6）SentiLR

Ke等人[85]将单词的词性、情感标签及句子的情感标签引入到预训练模型的编码信息中，将原有的MLM任务改造为LA-MLM（label-aware masked language model，标签感知的掩码语言模型）任务，该任务有两个子任务：①给定句子的情感标签，使模型预测词的信息（词、词性和词的情感标签）；②同时预测句子的情感标签和词的信息。该模型在多个情感分析的任务上有较好的表现，但在通用的任务上表现不稳定。

7）SKEP

Tian等人[86]在情感种子词和无监督的情感知识挖掘方法的基础上，提出了混合情感掩码策略，包括方面情感词对掩码、情感词掩码和通用词掩码。同时，结合掩码策略，提出了情感词预测（sentiment word prediction，SW）、词极性预测（word polarity prediction，WP）和方面情感词对预测（aspect-sentiment pair prediction，AP）。经过SKEP训练的预训练模型在多数情感分析任务上取得性能提升。

3.2.3 其他知识融合方法

由于预训练上下文编码器知识增强的方法种类较为丰富，除了融合显性和隐性知识的方法外，本文还列举了以下两种较为特别的类型，分别是知识集成模型和语言与知识联合模型。

1）知识集成模型

大多数的知识增强的上下文编码器都可以看作是对预训练上下文编码器在某些知识上的多任务学习，由于训练目标的多样性，当需要多种知识进行融合时，往往会发生知识遗忘问题，因此Wang等人[87]提出K-Adapter模型来解决知识集成问题。该模型通过不同的学习器（Adapter[102]）分别学习事实知识（factual knowledge）和语法知识（linguistic knowledge），在应用于下游任务时，将学习器和模型的特征进行拼接，使不同的知识都可以应用到任务中，是一种多类型知识集成的方法。

2）语言与知识联合模型

语言模型的输入以文本序列为主，而知识模型（或实体嵌入）的输入则是以三元组或知识子图为主（knowledge sub-graph）。在以往的工作中，实体嵌入（如TransE[92]、TuckER[95]等）模型通常是浅层且静态的，在一定程度上影响了实体嵌入的表示能力。因此，部分最新的研究考虑通过将预训练上下文编码器与知识模型相结合，实现上下文化的语言与知识联合嵌入。

KEPLER[88]模型借鉴了TransE中的方法，在预训练的过程中增加了对知识进行编码的任务，知识编码任务的损失函数如式（4）所示。

$$\mathcal{L}_{\mathrm{KE}}=-\log\sigma(\gamma-d_r(h,t))-\sum\frac{1}{n}\log\sigma(d_r(h'_i,t'_i)-\gamma) \tag{4}$$

对于给定的知识图谱三元组头实体、关系和尾实体(h,r,t)，其中，h和t是以知识库中对实体描述的句子为上下文，实体经过模型编码后的[CLS]位置的输出，(h'_i,r'_i,t'_i)为负样本，γ为超参数，d_r是知识嵌入分数的计算方法。KEPLER初始化为$\mathrm{RoBERTa}_{\mathrm{BASE}}$，但仅用该目标去优化模型，会发生灾难性遗忘问题[103]，因此在训练过程将其与预训练的MLM任务相结合。

Sun 等人[89]认为文本序列可以看作是全连接的词图，因此在 CoLAKE 模型中提出词-知识图(word-knowledge graph，WK graph)的概念。CoLAKE 仍将 Transformer 作为主干网络，将文本、实体和关系作为模型的输入，并利用类别嵌入对三者进行区分，同时模型也采用了与 K-BERT 相似的软位置编码，并利用遮蔽注意力来控制信息流。模型的训练目标包括三类掩码策略：词节点掩码(masking word nodes，MWN)、实体节点掩码(masking entity nodes，MEN)和关系节点掩码(masking relation nodes，MRN)。

KEPLER[88]和 CoLAKE[89]在知识相关的任务上有着较好的表现，但在一些通用任务上却存在一定的不稳定和效果下降，可以认为是知识图谱中缺少一般文本中的词语共现、语法习惯等信息，因此训练知识模型所需要的目标在一定程度上削弱了语言模型的泛化能力。

3.3　通用任务效果对比

大部分知识增强的预训练上下文编码器在结构上是任务通用的，虽然部分模型在引入的知识上有一定侧重，但其在通用领域任务上效果仍然具有一定的可比性。本节对比了部分模型在通用评测任务 GLUE(general language understanding evaluation)①上的效果，由于在不同文献中模型的参数设置不尽相同，因此只对比其与基线模型的效果差别，结果如表 2 所示。

表 2　模型在通用领域评测任务 GLUE 上的效果对比

模型名称	基线模型	MNLI (392k)	QQP (363k)	QNLI (104k)	SST-2 (67k)	CoLA (8.5k)	STS-B (5.7k)	MRPC (3.5k)	RTE (2.5k)
ERNIE(THU)[74]	$\text{BERT}_{\text{BASE}}$	↓	—	N/A	—	↑	↓	↓	↑
$\text{SenseBERT}_{\text{BASE}}$[77]	$\text{BERT}_{\text{BASE}}$	—	↓	↑	↓	↑	↓	↑	↓
LIBERT[78]	$\text{BERT}_{\text{BASE}}$	↑	↑	↑	↑	↑	↑	↑	↑
$\text{SemBERT}_{\text{BASE}}$[84]	$\text{BERT}_{\text{BASE}}$	↑	—	N/A	—	↑	↑	↓	↑
$\text{SemBERT}_{\text{LARGE}}$[84]	$\text{BERT}_{\text{LARGE}}$	↑	↑	↑	↓	↑	↑	↑	↑ ↑
KEPLER[88]	$\text{RoBERTa}_{\text{BASE}}$	↓	↑	—	↓	↓	↓	↑	↑ ↑
CoLAKE[89]	$\text{RoBERTa}_{\text{BASE}}$	↓	↑	↓	↓	↓	↓	↑	↓

注：“↑”和“↓”表示模型较基线模型效果提升和下降；—表示模型与基线模型效果持平；“↑ ↑”和“↓ ↓”表示模型较基线模型大幅提升和下降。

可以看出，外部知识的引入可能会对模型在通用评测任务上的效果有一定影响，但除个别任务上效果影响较大外，在大部分任务上都与基线模型持平或提升。

对于模型在部分任务上性能下降的原因，可以认为是引入的知识与任务无关[74,77]，与知识相关的训练目标在一定程度上影响了原有的训练目标。此外，引入不同类型的知识对模型在不同规模的任务数据集的效果影响也不同，例如 ERNIE[74]和 KEPLER[88]在小规模任务数据集上效果不稳定，而 SemBERT 则在小规模数据集上较基线模型有显著的性能提升。

4　总结与展望

当前预训练上下文编码器已逐渐取代预训练词嵌入，成为语言模型的主流，但预训练上下文编码器的知识增强仍然存在以下值得研究的问题：

1）知识增强与数据增强的比较

经过知识增强的预训练上下文编码器相较于数据增强的方法，是否能取得相当甚至更好的效果，仍然存在疑问。知识增强的方法可以在缺少相关数据的情况下辅助提升模型的泛化能力，特别是在应对特殊领域任务时，效果最为明显，但通过领域语料数据增量训练预训练模型仍然是当前主流的解决方案，并且在医学、司法、金融等专业领域，领域预训练模型在处理领域文本任务时已取得了显著的效果[104-105]。同时，由于不需要改变原有模型的结构和使用方法，领域预训练模型用起来更加便捷。

因此，需要对知识增强的预训练模型在特定领

① https://gluebenchmark.com/leaderboard

域中的效果做进一步的验证，使得知识驱动的语言模型在特定领域得到更加充分的应用。

2）知识增强的模式

如前文所述，当前知识增强的基本模式有两类：一类是在预训练阶段引入知识，另一类是在任务相关阶段引入知识，二者各有优势和缺点。

在预训练阶段引入知识，模型可以一次性将大量的知识融入到预训练模型中，其获得的知识是全局性的，需要消耗一定的预训练资源，在处理下游任务时无须再次调用知识库，在使用上更加便捷。但该模式的模型获得的知识仍是模糊的、非实时的，当出现新的知识时（如“新冠肺炎”相关知识），模型的可扩展性和可控性较差。

在任务相关阶段引入知识，需要根据任务文本对知识库进行检索，其获得的知识是局部的，引入模型中的知识是实时并且相对可控的，但模型的效果却依赖于知识检索与标注的算法和工具，容易造成误差累积或难以泛化等问题。

以上两种模式各有利弊，下一步的研究重点首先应研究更高效的知识检索方法，并对何时引入知识、何处引入知识、引入什么样的知识、引入多少知识等问题进行探索；其次，可以尝试两种模式结合的模型。

3）预训练上下文编码器与记忆网络结合

预训练上下文编码器模型大多采用多层Transformer[13]结构，当前，部分方法选择在模型的最后一层前引入知识[74-75,79]，因此可以通过模型的自注意力过滤掉引入知识中的“噪声”。由于采用了位置编码和自注意力机制，当需要引入不定长度和数量的知识文本时（例如长篇的解释或多个三元组），难以在不破坏原有模型结构的前提下将知识以“记忆”的方法引入到模型的输入中。

在下一步的研究中可以考虑结合Transformer-XL[106]、MemTransformer[107]等结构的预训模型进行知识增强，以此来编码不定长的外部知识。

参考文献

[1] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China (Technological Sciences), 2020,0:1872-1897.

[2] Almeida F,Xexéo G. Word embeddings: A survey[J]. arXiv preprint arXiv:1901.09069, 2019.

[3] Camacho Collados J, Pilehvar M T. From word to sense embeddings: A survey on vector representations of meaning[J]. Journal of Artificial Intelligence Research, 2018, 63: 743-788.

[4] 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(03): 162-173.

[5] Goldberg Y. A primer on neural network models for natural language processing[J]. Journal of Artificial Intelligence Research, 2016, 57: 345-420.

[6] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[7] Mikolov T, Chen K,Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[8] Pennington J,Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[9] Joulin A, Grave É, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[C]//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 427-431.

[10] Mccann B, Bradbury J, Xiong Caiming, et al. Learned in translation: Contextualized word vectors[C]//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 6294-6305.

[11] Peters M, Neumann M,Iyyer M, et al. Deep contextualized word representations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 2227-2237.

[12] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 4171-4186.

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference or Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 5998-6008.

[14] Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 3651-3657.

[15] Liu N F, Gardner M, Belinkov Y, et al. Linguistic knowledge and transferability of contextual represen-

tations[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg：ACL，2019：1073-1094.

[16] Wiedemann G，Remus S，Chawla A，et al. Does-BERT make any sense？Interpretable word sense disambiguation with contextualized embeddings[J]. arXiv preprint arXiv：1909. 10430，2019.

[17] Ethayarajh K. How contextual are contextualized word representations？Comparing the geometry of BERT，ELMo，and GPT-2 embeddings[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：55-65.

[18] Liu Y，Ott M，Goyal N，et al.RoBERTa：A robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv：1907. 11692，2019.

[19] Petroni F，Rocktäschel T，Riedel S，et al. Language models as knowledge bases？[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：2463-2473.

[20] Raffel C，Shazeer N，Roberts A，et al. Exploring the limits of transfer learning with a unified text-to-text Transformer[J]. arXiv preprint arXiv：1910.10683，2019.

[21] Kwiatkowski T，Palomaki J，Redfield O，et al. Natural questions：A benchmark for question answering research[J]. Transactions of the Association for Computational Linguistics，2019，7：453-466.

[22] Berant J，Chou A，Frostig R，et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2013：1533-1544.

[23] Joshi M，Choi E，Weld D S，et al. TriviaQA：A large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2017：1601-1611.

[24] Kassner N，Schütze H. Negated LAMA：Birds cannot fly[J]. arXiv preprint arXiv：1911.03343，2019.

[25] Poerner N，Waltinger U，Schütze H. BERT is not a knowledge base (yet)：Factual knowledge vs. Name-based reasoning in unsupervised QA[J]. arXiv preprint arXiv：1911.03681，2019.

[26] Niven T，Kao H-Y. Probing neural network comprehension of natural language arguments[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2019：4658-4664.

[27] Habernal I，Wachsmuth H，Gurevych I，et al. The argument reasoning comprehension task：Identification and reconstruction of implicit warrants[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg：ACL，2018：1930-1940.

[28] Kovaleva O，Romanov A，Rogers A，et al. Revealing the dark secrets of BERT[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：4356-4365.

[29] Searle J R. Minds，brains，and programs[J]. Behavioral and Brain Sciences，1980，3(3)：417-424.

[30] Talmor A，Elazar Y，Goldberg Y，et al.oLMpics—On what language model pre-training captures[J]. arXiv preprint arXiv：1912.13283，2019.

[31] 郁振华.波兰尼的默会认识论[J]. 自然辩证法研究，2001，(08)：5-10.

[32] Polanyi M. Knowing and being[J]. Mind，1961：458-470.

[33] Polanyi M. The study of man[M]. Chicago：University of Chicago Press，1959.

[34] Davies J，Studer R，Warren P. Semantic Web technologies：trends and research in ontology-based systems[M]. Hoboken，NJ：John Wiley and Sons，2006.

[35] Ehrlinger L，Wöß W. Towards a definition of knowledge graphs[C]//Proceedings of the Conference of SEMANTiCS. Vienna：Semantic Web Company，2016.

[36] Miller G A. WordNet：A lexical database for English[J]. Communications of the ACM，1995，38(11)：39-41.

[37] Navigli R，Ponzetto S P. BabelNet：The automatic construction，evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial Intelligence，2012，193：217-250.

[38] Chen X，Liu Z，Sun M. A unified model for word sense representation and disambiguation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2014：1025-1035.

[39] Chen T，Xu R，He Y，et al. A gloss composition and context clustering based distributed word sense representation model[J]. Entropy，2015，17(9)：6007-6024.

[40] Chen T，Xu R，He Y，et al. Improving distributed representation of word sense via wordnet gloss com-

position and context clustering[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Conference on Natural Language Processing. Stroudsburg: ACL, 2015: 15-20.

[41] Neelakantan A, Shankar J, Passos A, et al. Efficient non-parametric estimation of multiple embeddings per word in vector space[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1059-1069.

[42] Rothe S, Schütze H. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Conference on Natural Language Processing. Stroudsburg: ACL, 2015: 1793-1803.

[43] Pilehvar M T, Collier N. De-Conflated semantic representations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2016: 1680-1690.

[44] Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2015: 1606-1615.

[45] Mrkšić N, Vulić I, Séaghdha D Ó, et al. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints[J]. Transactions of the association for Computational Linguistics, 2017, 5: 309-324.

[46] Speer R, Lowry-Duda J. ConceptNet at SemEval-2017 Task 2: Extending word embeddings with multilingual relational knowledge[C]//Proceedings of the 11th Workshop on Semantic Evaluation. Stroudsburg: ACL, 2017: 85-89.

[47] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the Association for Computational Linguistics, 2015, 3: 211-225.

[48] Speer R, Chin J. An ensemble method to produce high-quality word embeddings[J].arXiv preprint arXiv:1604.01692, 2016.

[49] **孙茂松,陈新雄. 借重于人工知识库的词和义项的向量表示: 以** HowNet **为例**[J]. **中文信息学报**, 2016, 30(6): 1-6.

[50] Mancini M, Camacho-Collados J, Iacobacci I, et al. Embedding words and senses together via joint knowledge-enhanced training[C]//Proceedings of the 21st Conference on Computational Natural Language Learning. Stroudsburg: ACL, 2017: 100-111.

[51] Niu Y, Xie R, Liu Z, et al. Improved word representation learning with sememes[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 2049-2058.

[52] Zeng X, Yang C, Tu C, et al. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 5650-5657.

[53] Gu Y, Yan J, Zhu H, et al. Language modeling with sparse product of sememe experts[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4642-4651.

[54] Chen Q, Zhu X, Ling Z, et al. Neural natural language inference models enhanced with external knowledge[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2018: 2406-2417.

[55] Zou Y, Gui T, Zhang Q, et al. A lexicon-based supervised attention model for neural sentiment analysis [C]//Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: ACL, 2018: 868-877.

[56] Ghazvininejad M, Brockett C, Chang M W, et al. A knowledge-grounded neural conversation model[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 5110-5117.

[57] Parthasarathi P, Pineau J. Extending neural generative conversational model using external knowledge sources[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 690-695.

[58] Carlson A, Betteridge J, Kisiel B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2010: 1306-1313.

[59] Dinan E, Roller S, Shuster K, et al. Wizard of Wikipedia: knowledge-powered conversational agents [C]//Proceedings of the International Conference on Learning Representations. San Diego, CA: ICLR, 2018.

[60] Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 1870-1879.

[61] Mazare P E, Humeau S, Raison M, et al. Training millions of personalized dialogue agents[C]//Pro-

ceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2018：2775-2779.

[62] Meng C，Ren P，Chen Z，et al. RefNet：A reference-aware network for background based conversation[J]. arXiv preprint arXiv:1908.06449，2019.

[63] Yang B，Mitchell T. Leveraging knowledge bases in LSTMs for improving machine reading[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2017：1436-1446.

[64] Yang B，Yih W T，He Xiaodong，et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575，2014.

[65] Young T，Cambria E，Chaturvedi I，et al. Augmenting end-to-end dialogue systems with commonsense knowledge[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Menlo Park，CA：AAAI，2018.

[66] Mihaylov T，Frank A. Knowledgeable reader：Enhancing cloze-Style reading comprehension with external commonsense knowledge[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2018：821-832.

[67] Miller A，Fisch A，Dodge J，et al. Key-value memory networks for directly reading documents[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2016：1400-1409.

[68] Zhong P，Wang D，Miao C. Knowledge-enriched Transformer for emotion detection in textual conversations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：165-176.

[69] Zhou H，Young T，Huang M，et al. Commonsense knowledge aware conversation generation with graph attention[C]//Proceedings of the 27th International Conference on Artificial Intelligence. San Francisco，CA：Morgan Kaufmann，2018：4623-4629.

[70] **李强，黄辉，周沁，等. 模板驱动的神经机器翻译**[J]. **计算机学报**，2019，42(03)：116-131.

[71] Huang L，Sun C，Qiu X，et al. GlossBERT：BERT for word sense disambiguation with gloss knowledge[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：3500-3505.

[72] Guu K，Lee K，Tung Z，et al. Realm：Retrieval-augmented language model pre-training[J]. arXiv preprint arXiv:2002.08909，2020.

[73] Yang A，Wang Q，Liu J，et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2019：2346-2357.

[74] Zhang Z，Han X，Liu Z，et al. ERNIE：Enhanced language representation with informative entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg：ACL，2019：1441-1451.

[75] Peters M E，Neumann M，Logan R，et al. Knowledge enhanced contextual word representations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg：ACL，2019：43-54.

[76] He B，Zhou D，Xiao J，et al. Integrating graph contextualized knowledge into pre-trained language models[J]. arXiv preprint arXiv:1912.00147，2019.

[77] Levine Y，Lenz B，Dagan O，et al. SenseBERT：Driving some sense into BERT[J]. arXiv preprint arXiv:1908.05646，2019.

[78] Lauscher A，Vulić I，Ponti E M，et al. Specializing unsupervised pretraining models for word-level semantic similarity[J]. arXiv preprint arXiv：1909.02339，2020.

[79] 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. Menlo Park，CA：AAAI. 2020：2901-2908.

[80] Sun Y，Wang S，Li Y，et al. ERNIE：Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223，2019.

[81] Cui Y M，Che W X，Liu T，et al. Pre-training with whole word masking for Chinese BERT[J].arXiv preprint arXiv:1906.08101，2019.

[82] Xiong W，Du J，Wang W，et al. Pretrained encyclopedia：Weakly supervised knowledge-pretrained language model[J]. arXiv preprint arXiv:1912.09637，2019.

[83] Yamada I，Asai A，Shindo H，et al. LUKE：Deep contextualized entity representations with entity-aware self-attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2020：6442-6454.

[84] Zhang Z，Wu Y，Zhao H，et al. Semantics-aware BERT for language understanding[J]. arXiv preprint

arXiv:1909.02209, 2019.

[85] Ke P, Ji H, Liu S, et al. SentiLR: Linguistic knowledge enhanced language representation for sentiment analysis[J]. arXiv preprint arXiv:1911.02493, 2019.

[86] Tian H, Gao C, Xiao X, et al. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis[J]. arXiv preprint arXiv:2005.05635, 2020.

[87] Wang R, Tang D, Duan N, et al. K-adapter: Infusing knowledge into pre-trained models with adapters [J]. arXiv preprint arXiv:2002.01808, 2020.

[88] Wang X, Gao T, Zhu Z, et al. KEPLER: A unified model for knowledge embedding and pre-trained language representation[J]. arXiv preprint arXiv: 1911.06136, 2020.

[89] Sun T,Shao Y, Qiu X, et al. CoLAKE: Contextualized language and knowledge embedding[C]//Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: ACL, 2020: 3660-3670.

[90] Clark K, Luong M T, Le Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators[C]//Proceedings of the International Conference on Learning Representations. San Diego, CA: ICLR, 2019.

[91] Lan Z, Chen M, Goodman S, et al. ALBERT: A Lite BERT for self-supervised learning of language representations[C]//Proceedings of the International Conference on Learning Representations. San Diego, CA: ICLR, 2020.

[92] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 26th International Conference or Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795.

[93] Ferragina P, Scaiella U. Tagme: on-the-fly annotation of short text fragments (by wikipediaentities) [C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010: 1625-1628.

[94] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference or Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 3111-3119.

[95] Balazevic I, Allen C, Hospedales T. TuckER: Tensor factorization for knowledge graph completion [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Conference on Natural Language Processing. Stroudsburg: ACL, 2019: 5188-5197.

[96] Spitkovsky V I, Chang A X. A Cross-lingual dictionary for English Wikipedia concepts[C]//Proceedings of the 8th International Conference on Language Resources and Evaluation. Stroudsburg: ACL, 2012: 3168-3175.

[97] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2011: 782-792.

[98] Vulić I. Injecting lexical contrast into word vectors by guiding vector space specialisation[C]//Proceedings of the 3rd Workshop on Representation Learning for NLP. Stroudsburg: ACL, 2018: 137-143.

[99] Sun Y, Wang S, Li Y, et al. ERNIE 2.0: A continual pre-training framework for language understanding [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2020, 34 (5):8968-8975.

[100] Che W, Li Z, Liu T. LTP: A Chinese language technology platform[C]//Proceedings of the COLING. Stroudsburg: ACL, 2010: 13-16.

[101] Wei J, Ren X, Li X, et al. NEZHA: Neural contextualized representation forchinese language understanding[J]. arXiv preprint arXiv: 1909.00204, 2019.

[102] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]//Proceedings of the International Conference on Machine Learning. New York: ACM, 2019: 2790-2799.

[103] Mccloskey M, Cohen N J. Catastrophic interference in connectionist networks: the sequential learning problem[J]. The Psychology of Learning and Motivation, 1989, 24: 109-165.

[104] Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 8342-8360.

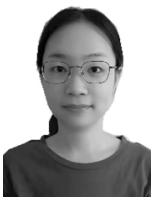
[105] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36 (4): 1234-1240.

[106] Dai Z, Yang Z, Yang Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 2978-2988.

[107] Burtsev M S, Sapunov G V. Memory Transformer [J]. arXiv preprint arXiv:2006.11527, 2020.

孙毅(1993—)，博士研究生，主要研究领域为自然语言处理、信息检索。
E-mail：sunyi_lgdx@sina.com

裘杭萍(1965—)，通信作者，博士，教授，主要研究领域为系统工程、信息检索。
E-mail：qiuhp_zy@163.com



郑雨(1994—)，硕士，主要研究领域为信息检索、自然语言处理、智能软件测试。
E-mail：zhengyu87@outlook.com

第一届中国情感计算大会暨情感计算专委会(筹)工作会议顺利召开

2021 年 7 月 10—11 日，第一届中国情感计算大会(First Chinese Conference on Affective Computing，CCAC 2021)暨中国中文信息学会情感计算专委会(筹)工作会议在北京会议中心隆重举行。本次会议由中国中文信息学会主办，学会情感计算专委会(筹)协办、清华大学承办。大会由大连理工大学林鸿飞教授、美国伊利诺伊大学芝加哥分校刘兵教授担任大会共同主席，哈尔滨工业大学(深圳)徐睿峰教授、清华大学贾珈长聘副教授担任程序委员会共同主席，清华大学黄民烈长聘副教授担任组织委员会主席，复旦大学魏忠钰副教授担任宣传主席，国际关系学院李斌阳副教授担任赞助主席。来自全国各高校、研究机构和企事业单位的共 400 余位代表参加了本次会议。

7 月 11 日第一届中国情感计算大会正式开幕。开幕式由大会程序委员会共同主席徐睿峰教授主持。大会共同主席林鸿飞教授代表大会组委会进行致辞。他介绍了会议整体的筹备和参会情况，对各位代表齐聚北京表示欢迎，并对大会赞助商表示感谢。中国中文信息学会学会名誉理事长李生教授在致辞中表示情感计算后继有人，值得高兴。目前，大数据、人工智能、脑科学、生物科学、社会学等多学科出现相互启发、深入交叉融合的趋势。今年中国中文信息学会成立了情感计算专委会，创办了中国情感计算大会，为相关的学术交流提供了一个平台。希望通过这次大会，大家取长补短，互通有无，深入交流。学会原副理事长、欧洲科学院院士、清华大学孙茂松教授代表承办方欢迎大家参加本次会议。他表示，情感是人类的高级认知和表达，在多个领域都具有独特的研究价值。他期待早日看到更多可以实用的情感计算系统。最后，他代表清华大学预祝此次大会圆满成功。学会副理事长兼秘书长孙乐研究员代表学会致辞，他表示学会高度关心和支持情感计算专委会(筹)的发展，希望专委会做好各方面工作，促进领域内专家学者的合作和交通。情感计算专委会(筹)主任秦兵教授向与会代表介绍了专委会的发展情况和组织情况，她鼓励更多学者加入专委会，以及组织更多的专业组和工作组，共同做好一个有情感、有温度、学科有交叉的专委会。她特别感谢了承办单位清华大学对本次大会的支持。学会副理事长、拓尔思信息技术公司施水才总裁也出席了本次开幕式。

大会的闭幕式由徐睿峰教授主持，苏州大学王中卿老师代表第二届情感计算大会承办方苏大自然语言处理实验室介绍了承办方的基本情况，并热情邀请各位代表参与 CCAC 2022。最后，秦兵老师代表专委会宣布大会圆满结束，期待大家 2022 年在苏州相聚。