知识增强预训练模型

Knowledge-Enhanced Pre-Trained Models

王海峰/WANG Haifeng，孙宇/SUN Yu，吴华/WU Hua
（北京百度网讯科技有限公司，中国 北京 100193）
(Beijing Baidu Netcom Science and Technology Co., Ltd., Beijing 100193, China)

DOI：10.12142/ZTETJ.202202004
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20220408.1423.006.html
网络出版日期：2022-04-08
收稿日期：2022-02-16

摘要：预训练模型主要从海量未标注、无结构化的数据中学习，但缺少外部知识指导，存在模型学习效率不高、模型效果不佳和知识推理能力受限等不足。如何在预训练模型中引入语言知识、世界知识等外部知识，提升模型效果以及知识记忆和推理能力是一个难题。本文从不同类型知识的引入、融合知识的方法、缓解知识遗忘的方法等角度，介绍知识增强预训练模型的发展，并以知识增强预训练模型百度文心为例，详细探讨知识增强预训练模型的原理和应用。

关键词：自然语言处理；预训练模型；知识增强

Abstract: Pre-trained models can automatically learn from massive unmarked and unstructured data. Nevertheless, the lack of guidance from external knowledge has dramatically hindered the learning efficiency, model effect, and reasoning capacity. There are still challenges in incorporating external supervision such as linguistics and world knowledge to improve pre-trained models' ability of knowledge memorization and reasoning. This paper provides a comprehensive review of knowledge-enhanced pre-trained models from various perspectives, such as multi-source knowledge incorporation, knowledge fusion, and knowledge forgetting alleviation. Here we take Baidu ERNIE as an example to describe the principles and applications of knowledge-enhanced pre-trained models.

Keywords: natural language processing; pre-trained model; knowledge-enhanced

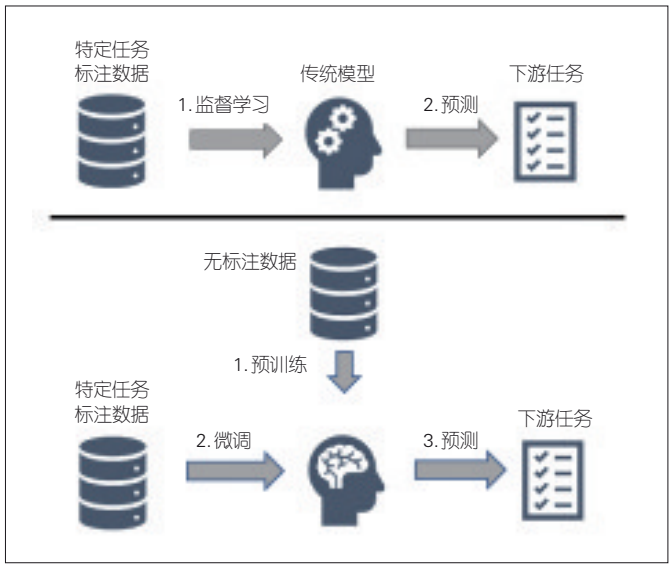
自然语言处理中的预训练模型与语言模型的建立密切相关。语言模型是自然语言处理的一个重要分支。早期的语言模型能够对由单词组成的文本序列进行概率建模[1-2]，并计算句子的联合概率。该模型技术被广泛应用于自然语言处理任务中，例如语音识别[3]、机器翻译[4]等。

2003年，随着深度学习技术的发展，Y. BENGIO等提出神经网络语言模型NNLM[5]。该模型被用来学习词的分布式表示以解决词表示的维数灾难问题。2013年，词表示训练技术Word2Vec[6]被提出。该技术可使用词的上下文来对当前词进行建模，从而学习单词的分布式向量表示。随后，一系列词表示技术如雨后春笋般涌现，例如基于词汇共现矩阵的GloVe[7]、基于字符级别N-Gram的FastText[8]等。词表示技术的提出是深度学习在自然语言处理方向应用的一座里程碑。这种技术极大地加速了自然语言处理领域的发展进程。

由于Word2Vec[6]词表示技术仅能将语言中的词语映射到一个静态的、与上下文无关的语义表示空间上，因此该技术无法解决语言中的一词多义问题。2018年，ELMo模型[9]采用了双向长短期记忆（LSTM）网络对文本序列的上下文进行建模。该模型能够将双向语言模型中的不同层表示进行融合，并计算上下文相关的词表示，在一定程度上解决了一词多义问题。紧接着，BERT模型[10]使用双向Transformer网络[11]对文本序列进行建模，并采用预训练-微调方法一举刷新众多自然语言理解任务的基准纪录。预训练模型技术的成熟进一步推动了自然语言处理的发展。

与传统监督学习方法不同，基于自监督学习方法的预训练-微调首先对大规模无标注数据进行学习，然后再对小规模任务标注数据进行微调，如图1所示。由于能同时对未标注文本和标注文本进行学习，预训练-微调方法取得了远超传统监督学习的效果，并且显著缩小了任务标注数据的规模[10,12]。因此，预训练-微调方法逐渐成为自然语言处理领域的应用范式。近期，基于超大规模预训练模型的预训练-提示方法[13]取得了能够与预训练-微调方法相媲美的效果，并逐渐成为自然语言处理领域的又一范式。该方法可将下游任务改造为自然语言表达形式，使下游任务的建模形式更接近预训练模型的学习过程，从而挖掘出预训练模型强大的零样本和小样本学习能力。

得益于深度学习技术和硬件算力的飞速发展，以BERT[10]、GPT-3[12]、ERNIE 3.0[14]为代表的预训练语言模型



▲图1 传统监督学习(上)与预训练-微调(下)的对比

在自然语言理解、语言生成、机器翻译、人机对话等领域取得了突破性进展。预训练模型的出现使得人们对自然语言处理领域的研究重点从过去的结构工程转移到目标工程上，即从设计不同的网络结构并引入相应的归纳偏置，转移到基于统一的Transformer模型来设计启发式的预训练目标。预训练模型凭借自监督学习方法和预训练-微调应用方法，已逐步占据自然语言处理领域的主导地位。

当前的预训练模型主要依赖大量无结构化数据的学习。由于缺少外部知识指导，这些模型存在学习效率不高、模型效果不佳和知识推理能力受限等问题。因此，如何使用知识来增强预训练模型的表示能力，是预训练模型研究和应用的难点之一。目前，主流的知识增强预训练模型主要分为两类。一类模型可通过弱监督方法，对文本中蕴含的知识进行标注，然后设计知识类预训练任务，以便对文本中的知识进行学习。例如，ERNIE 1.0[15]通过对数据中的短语和实体进行标注并掩码，来学习文本中的知识。文献[16]对实体知识进行替换，使语言模型能够根据上下文信息对知识图谱中的实体和关系进行推断，从而加强对文本序列知识的学习。另一类模型可对构建好的结构化知识库和无结构化文本进行联合预训练学习，例如K-BERT[17]、CoLAKE[18]和ERNIE 3.0[14]。通过对结构化知识和海量无结构化数据的联合学习，知识增强的预训练模型可以很好地提升知识记忆能力和推理能力[14]。

1 3种知识增强预训练模型

根据融合知识的类型和作用，本文将预训练模型分为3类：融合语言知识的预训练模型、融合世界知识的预训练模型和融合领域知识的预训练模型。

1.1 融合语言知识的预训练模型

语言知识是理解自然语言的基础，主要包含词法知识、句法结构知识、语义知识等。预训练模型对语言知识的融合方法有两种：一种是通过自动标记无标注文本中的语言知识来指导预训练模型的学习，另外一种是融合人工构建的语言知识库。ERNIE-Gram[19]通过构建基于N-Gram的多粒度掩码语言模型，可同时学习N-Gram内部和N-Gram之间的语义关系，使模型能够同时捕获细粒度和粗粒度语言知识，显著提升了模型的语义表示能力。除了融合语言粒度知识外，也有工作研究如何学习句子中的语义关系。通过在预训练的过程中对指代消解进行建模，CorefBERT[20]增强了模型对语义知识的学习能力。其中，“指代”是自然语言表达中的常见现象。基于在一段文本中多次出现的命名实体是同一个事物的假设，CorefBERT提出提及指代预测算法。通过预测文本中被掩盖的、重复出现的命名实体，该算法提升了模型对指代关系的建模能力。

上述方法主要对无标注数据中蕴含的人类知识进行标注，让模型通过学习标注信息来融合语言知识。此外，也有研究将人工构建的语言知识库融合到预训练模型中。其中，WordNet[21]和HowNet[22]是具有代表性的语言知识库。这些知识库含有丰富的语言知识。以WordNet为例，它将不同词性的单词各自组成一个同义词集合。每个同义词集合各表示一个基本的语义概念。WordNet利用语义关系将这些集合连接成网络。其中，每个词语均有对应的解释和例句。Sense-BERT[23]融合了WordNet中的超义等概念知识。通过还原被掩盖的词并预测其对应的超义，该模型可以显式学习词语在给定语境下的语义信息。SenseBERT在词义消歧等任务上的效果取得了显著提升。LIBERT[24]利用WordNet中词语与词语间的同义关系和上下位关系设计了词汇关系分类预训练任务过程，增强了预训练模型对语义信息的建模能力，在大部分自然语言处理任务上的效果均有提升。

1.2 融合世界知识的预训练模型

人类在认识世界的过程中产生了大量的世界知识。其中，部分知识可以利用实体以及实体之间的关系进行描述，比如“安徒生”创作了“《夜莺》”。研究者通过知识图谱[25]来表达这些世界知识。在知识图谱中，实体表示网络中的一个节点，实体间的关系则表示对应节点间的边。利用知识图谱存储世界知识，并让模型显式学习人类对世界的认

知，是融合世界知识的预训练模型采用的重要方法。KEPLER[26]将预训练上下文编码器与知识模型相结合，使得预训练模型不仅可以将图谱三元组中的事实知识更好地融合到模型中，而且还可以通过丰富的实体描述，有效地学习实体和关系的知识表示。不同于KEPLER，有的模型将语言和知识进行统一表示。CoLAKE[18]将文本序列视为一个全链接的词图，并以每个实体为锚点，将文本中实体所对应的知识图谱中的子图进行连接，以构成一个同时包含词语、实体和关系的词语–知识图。通过学习词语–知识图，模型能够同时融合训练语料中的语言知识和图谱中的世界知识。然而，CoLAKE主要侧重实体在知识图谱中的建模，却忽视了实体在训练语料中的表述。为此，ERNIE 3.0[14]提出知识图谱与文本平行预训练的方法，使用文本来表述知识。ERNIE 3.0突破了异构结构化知识表示与无结构文本表示难以统一建模的瓶颈。

1.3 融合领域知识的预训练模型

人工智能行业应用存在着丰富的、由众多行业专家积累的专业知识。当前的预训练模型主要依赖互联网数据进行训练。数据中缺乏行业相关的领域知识，导致预训练模型在专业领域的自然语言处理任务上的表现不佳。以医疗领域为例，CBLUE[27]的应用表明，通用预训练模型处理该类任务的效果差于人类。为了增强预训练模型在专业领域的应用效果，研究者们对如何将领域知识融入到预训练模型进行了探索。BioBERT[28]是一个生物医学领域的预训练模型。实验表明，在生物医学语料库上的预训练可以显著提高模型在生物医疗领域任务上的性能。针对领域知识的预训练方法，ERNIE–Health利用医疗实体掩码算法对专业术语等实体知识进行学习。同时，通过医疗问答匹配任务，该模型能对病状描述与医生专业治疗方案的对应关系进行学习，可获得医疗实体知识之间的内在联系，在包含医学信息抽取、医学术语归一化等中文医疗文本处理任务上的效果取得了显著提升。进一步地，结合世界知识和领域知识的学习方法，BERT–MK[29]基于医疗知识图谱的子图进行学习，提高了预训练模型在医疗领域任务上的应用效果。

为了充分地融合领域知识，以FLAN[30]、ExT5[31]和T0[32]为代表的模型分别收集了60、107、171个领域的任务数据，并针对每项任务设计了任务模板。将多种多样的任务转化为由文本至文本生成的统一格式，使模型在预训练阶段就能融合并使用多领域、多任务的知识，可显著提高模型的通用能力与泛化性能。PPT[33]延续了将多种任务通过模板转化为统一格式的方式，在预训练阶段就可对连续提示词进行领域知识的学习，提升了模型在训练样本匮乏的下游任务上的少样本迁移能力。

知识增强预训练模型通过融合多种类型的外部知识来显著提升自身性能。然而，在学习知识的过程中，模型通常存在知识遗忘问题，即在学习新的知识后会忘记之前学过的知识。因此，如何解决知识遗忘问题显得非常重要。为了避免知识遗忘，ERNIE 2.0[34]构建了持续预训练的框架。在该框架下，每当引入新任务时，该框架可在学习该任务的同时仍记住之前学过的知识。此外，K–ADAPTER[35]通过不同的适配器来学习世界知识和语言知识。在下游任务中，该方法能够将不同适配器产生的特征表示进行拼接，并生成同时具有各种知识的表示，从而将多种知识同时应用到任务中，有效解决了知识遗忘问题。

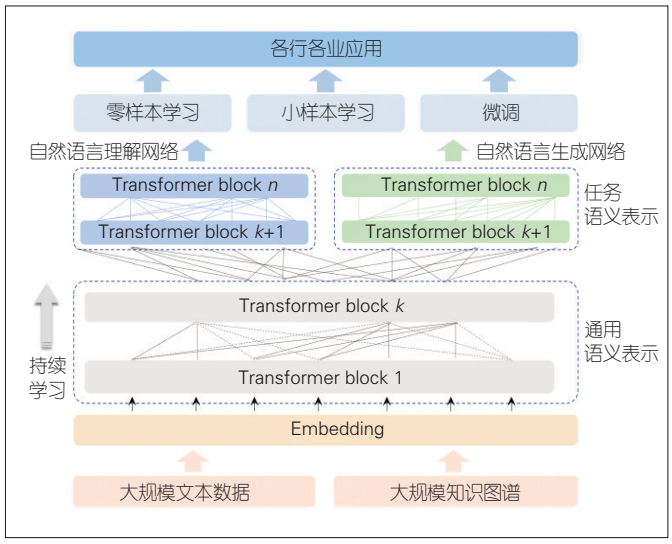
2 文心知识增强预训练模型

本文中，我们将以百度文心（ERNIE）知识增强预训练模型为例，详细阐述知识增强预训练模型的模型结构、知识融合方法，以及该模型在知识增强跨语言预训练模型、知识增强跨模态预训练模型上的扩展。文心是最早探索预训练模型融入知识的工作[15]之一，并在文献[14]和文献[34]等工作中逐步迭代。其中，最新的ERNIE 3.0 Titan模型[36]使用2 600亿个参数，在海量的未标注文本数据和大规模知识图谱中持续学习，突破了多源异构数据难以统一表示与学习的瓶颈，在60余项任务上的表现是最好的。

2.1 模型结构

文心使用了一种通用语义表示与任务语义表示相结合的模型框架，如图2所示。该框架融合了自编码和自回归等不同的任务语义表示网络。因此，文心既可以同时完成语言理解和语言生成任务，又能进行无标注数据的零样本学习和有标注数据的微调训练。该模型结构共包括两层：第1层是通用语义表示网络，该网络主要学习数据中的基础知识和通用知识；第2层是任务语义表示网络，该网络可基于通用语义表示来学习与任务相关的知识。不同任务语义表示网络可通过自编码结构或者自回归结构来实现。底层共享有助于这些任务语义表示网络实现交互和增强。在学习过程中，任务语义表示网络只学习对应类别的预训练任务，而通用语义表示网络则学习所有的预训练任务。

文心将Transformer[11]作为基础的模型结构，通过多层统一的自注意力机制，采用并行计算的方式来获得词与词之间的关系权重，并根据所得到的权重来生成每个词在整段语义单元的动态词表示。为了增强模型对长距离语义知识的建模



▲图2 文心模型结构

能力，文心引入了递归性记忆单元[37]，并在此基础上形成了一种增强记忆力机制[38]，使模型能够对超长文本进行建模。

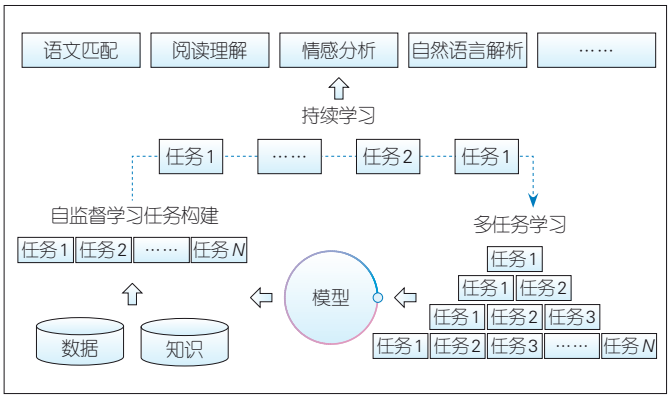
2.2 持续学习方法

ERNIE 2.0[34]拥有一种持续学习的预训练框架，可增量学习海量数据中的知识，持续提升语义理解效果。如图3所示，知识可通过预训练任务的形式加入训练框架。每当引入新的预训练任务时，该框架可在学习新任务的同时学习之前的任务。新任务与旧任务之间通过多任务进行学习可避免知识遗忘。基于该框架，模型可以快速学习词法、结构、语义层面的语言知识、实体–关系世界知识等。模型的通用能力可得到大幅提升。ERNIE 2.0将这种学习方式与传统的持续学习及多任务学习进行对比，结果证明了该方法的有效性。

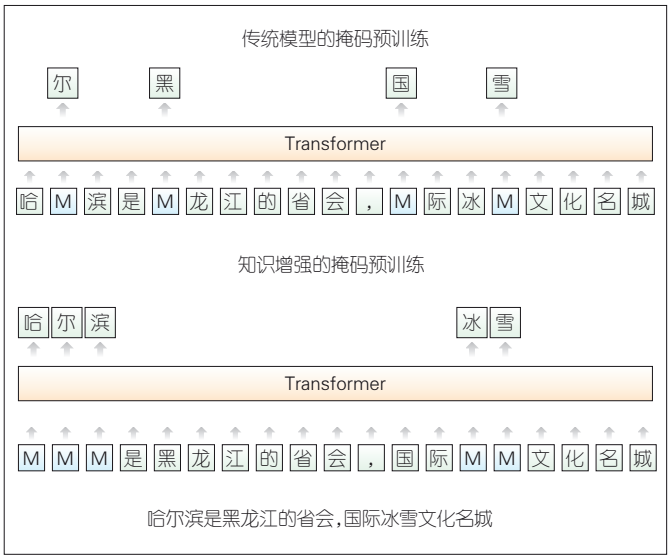
2.3 知识融合方法

2.3.1 语言知识融合方法

ERNIE 1.0[15]模型提出了知识增强的预训练方法，即知识掩码预训练方法。该模型通过对海量数据中的字、词、实体等不同语言单元和知识进行建模，来学习不同粒度语言知识的完整语义。图4给出了传统预训练模型和ERNIE 1.0学习方法的对比。在预测还原过程中，传统预训练模型通过诸如“哈尔滨”“黑龙江”等短距离固定记忆对被掩码的字进行还原，难以学习到“哈尔滨”“黑龙江”等命名实体的完整语义。而在ERNIE 1.0的学习过程中，只有学习到“哈尔滨”“黑龙江”等命名实体的关系，“哈尔滨”这一命名实体的属性才能正确预测被掩盖的知识。ERNIE 1.0本身可基



▲图3 文心模型中的持续学习语义理解框架



▲图4 文心语言知识学习方法

于字特征输入完成建模，在应用时不需要依赖其他信息，具有很强的通用性和可扩展性。例如，在对红色、绿色、蓝色等表示颜色的词语进行建模时，ERNIE 1.0通过相同字的语义组合可以学习词之间的语义关系。

在语义知识融合方面，短句中的连词往往准确地表示了它们的细分逻辑语义关系。例如，在“因为人们的滥砍乱伐，所以今年以来洪涝不断”中，“人们的滥砍乱伐”和“近年来洪涝不断”就是因果关系；“尽管风雨交加，但是同学们还是坚持按时到校上课”中的“风雨交加”和“同学们还是坚持按时到校上课”之间就是转折关系。为了能够实现短句间的逻辑关系建模，文心构建了逻辑关系知识：首先将具有逻辑关系的句子挖掘出来，然后再将句子中的连词去掉，最后让模型进行无监督的逻辑关系分类。

2.3.2 世界知识融合方法

ERNIE 3.0在引入蕴含丰富世界知识的大规模知识图谱

后，实现了海量无监督文本与大规模知识图谱的平行预训练。以图5为例，ERNIE 3.0在训练过程中会将文本端信息和知识端信息同时输入到模型中进行训练。知识端信息会输入图谱中的三元组。例如，“安徒生”“作品”“《夜莺》”三元组代表了《夜莺》是安徒生的作品这一世界知识。文本端就会使用三元组中的“安徒生”和“《夜莺》”在海量文本中检索出与之相关的句子。ERNIE 3.0在训练过程中使用联合掩码进行训练。训练过程主要包括两个方面：在知识端方面，由于知识图谱中的世界知识片段会被掩盖，模型需要通过文本中的信息对知识端被掩盖的信息进行推理；在文本端方面，由于无标注文本的语言知识片段也会被掩盖，模型需要通过图谱中的结构化信息对文本端被掩盖的信息进行还原。这种方式促进了结构化的知识和无结构文本之间的信息共享，大幅提升了模型对知识的记忆和推理能力。

与CoLAKE[18]、K-BERT[17]、KG-BART[39]、KnowBert[40]等融入知识图谱的工作原理不同，ERNIE 3.0利用知识图谱中三元组文本表述和对应的文本信息，在统一的空间同时对知识端和文本端进行平行学习。而先前的知识增强方法在融合知识与文本时使用了不同的编码结构，使得知识与文本只能在不同的表示空间中被学习。大部分研究工作只强调知识对文本的增强，却忽略文本对知识的作用，致使文本与知识的交互不充分。ERNIE 3.0增强了结构化知识与无结构文本间的双向交互，提升了模型对知识的理解与推理能力。

2.4 文心模型效果分析

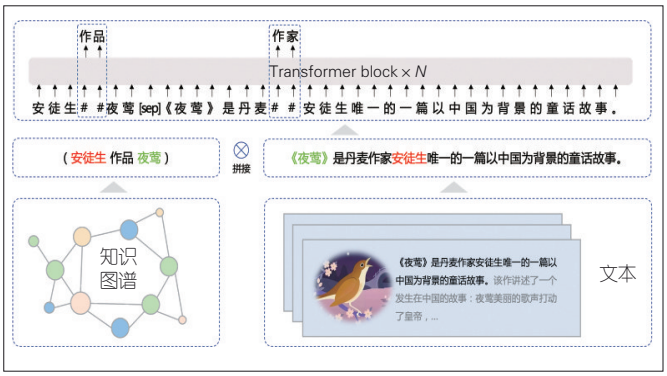
文心所使用的一系列知识增强方法显著提升了模型效果和学习效率，增强了知识推理能力。

知识增强预训练模型显著提升了下游任务效果。通过知识融合，相对于其他预训练模型，ERNIE 3.0[14]模型在包括情感分析、信息抽取、对话生成、数学计算、阅读理解等21类54个自然语言理解和生成数据集上的效果是最好的。表1表明，在语义匹配、文本摘要等任务上，只用3%的参数量，知识增强预训练模型就可以达到甚至超过百亿参数非知识增强预训练模型[41-42]的效果。同时，百亿参数的知识增强预训练模型效果可以得到进一步提升。

知识增强预训练模型的知识推理能力也得到了进一步提升。图6给出了ERNIE 3.0 Titan模型和GPT-3模型在知识问答数据集上的对比效果[36]。其中，ERNIE 3.0 Titan的准确率比GPT-3高8%。

2.5 知识增强预训练模型扩展

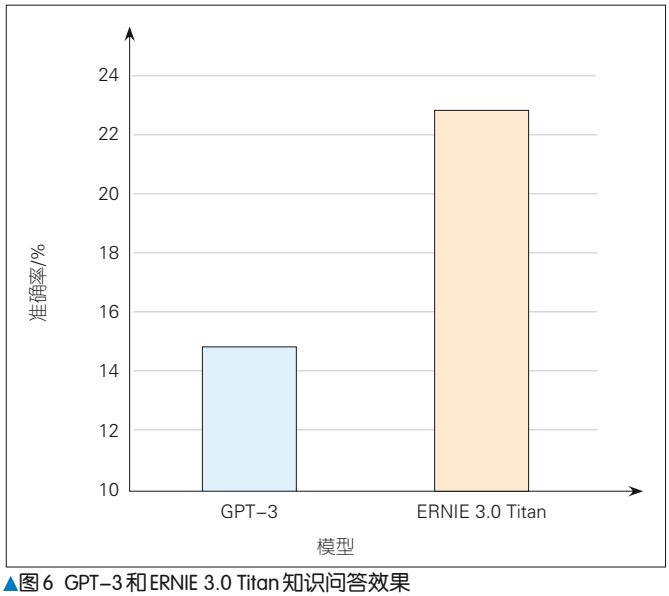
在单语言理解与生成预训练模型的基础上，为了融合更多维度的知识，文心进一步衍生出知识增强跨语言模型和知识增强跨模态模型。



▲图5 文心中的文本与知识平行预训练

▼表1 传统模型与知识增强模型效果对比

模型	是否知识增强	参数规模/亿	语义匹配准确率/%	文本摘要指标（Rouge-L）
mT5-XXL	否	110	88.3	34.8
CPM-2	否	130	89.2	35.9
文心	是	3	89.1	41.4
		100	90.4	48.4



▲图6 GPT-3和ERNIE 3.0 Titan知识问答效果

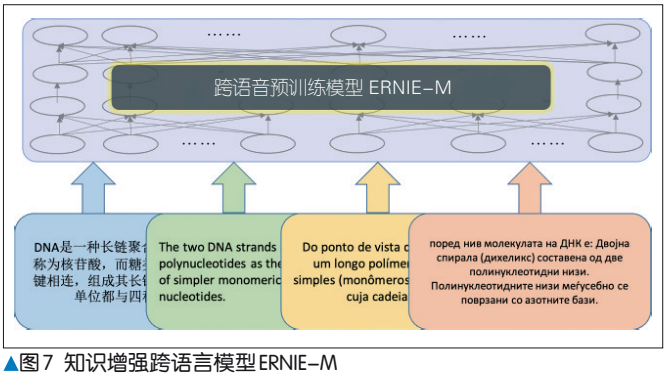
2.5.1 知识增强跨语言预训练模型

不同语言中的语料蕴含了不同地区的人们在历史发展过程中收集的不同知识。受限于语料的不完备性，模型从单一语言的语料中难以完全学到跨语言知识。因此，我们需要探索将多种语言数据中的知识进行融合的方法，以提升模型能力，解决单一语言数据的知识稀疏性问题。

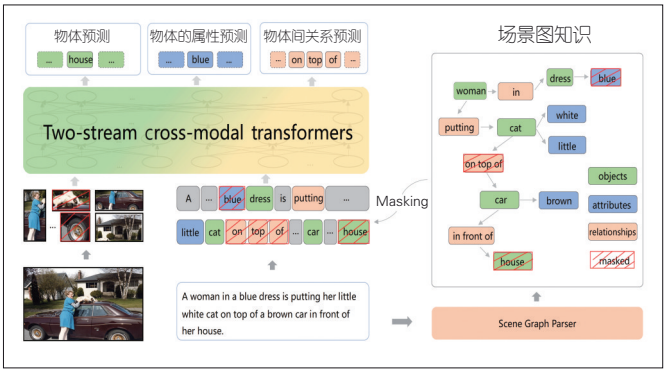
知识增强跨语言预训练模型[43]实现了从多种语言数据中进行跨语言知识学习的目标。在预训练过程中，ERNIE-M会使用统一的模型同时对海量未标注的多语言数据进行建模，从而统一学习跨语言知识和跨语言语义表示。如图7所示，对于DNA这一知识，不同语言的语料蕴含了不同的信息。因此，模型可以从不同语言中学到跨语言知识的不同侧面。在跨语言预训练模型使用某种语言的任务数据进行训练后，其他语言的相同任务无须进行进一步训练，即可实现跨语言迁移。这种跨语言迁移方式能够解决低资源语言任务数据稀疏性问题，有助于实现任务知识在不同语言间的迁移。从单语语料中学习多语间的隐式语义对齐知识的方法，能够突破双语平行语料规模对跨语言模型的限制。ERNIE-M对96种语言进行统一建模，并在5项跨语言任务中取得了最好的效果[43]。

2.5.2 知识增强跨模态模型

跨模态表示学习的目标是，通过对齐语料学习跨模态的通用联合表示，将各个模态之间的语义对齐信号融合到联合表示中，从而提升下游任务效果。目前的视觉-语言跨模态预训练方法，例如ViLBERT[44]等，在预训练过程中无法区分普通词和与场景相关的词，学到的联合表示也无法实现模态



▲图7 知识增强跨语言模型ERNIE-M



▲图8 跨模态知识增强模型ERNIE-ViL

间细粒度语义（如物体、物体属性、物体间关系）的对齐。

ERNIE-ViL[45]将包含细粒度语义信息的场景图先验知识融入视觉-语言跨模态预训练过程中，如图8所示。基于场景图的结构化知识，ERNIE-ViL创建物体预测、属性预测、关系预测3个预训练任务，在预训练过程中更加关注细粒度语义的跨模态对齐，从而可以学习到能够刻画更好跨模态语义对齐信息的联合表示，并提升自身在视觉问答、视觉常识推理、引用表达式理解、跨模态文本-图像检索等5个多模态典型任务上的应用效果。

3 知识增强预训练模型应用

随着预训练技术的快速发展，知识增强预训练模型有着非常广阔的应用场景，例如搜索引擎、推荐系统、智能创作、人机对话、文档分析、金融风控、智慧医疗等。这里，我们将从搜索引擎、人机对话、行业领域应用3个方面，详细阐述知识增强预训练模型的应用。

3.1 搜索引擎应用

搜索引擎通过对网页内容和用户查询请求进行分析和理解，让用户可以在海量的互联网数据中查询到所需的信息。通用的预训练模型很好地提升了搜索引擎效果，例如：谷歌在BERT问世一年之际宣布将预训练模型应用到搜索引擎中，并称BERT比以往任何技术都能更好地理解用户搜索意图；微软将Turing-NLG模型应用在必应搜索方案中，使得搜索引擎在搜索框内即可辅助用户完成查询词的输入；在中文搜索引擎中，百度将知识增强的文心模型运用到搜索引擎的不同检索阶段，包括端到端的大规模语义索引系统[46]、精细化语义相关性建模[47]、智能问答等。得益于基于大规模文本和大规模知识的自监督训练，文心模型可以帮助搜索引擎更加准确地理解网页内容和用户查询语句，从而提升搜索结果的准确性。传统的搜索引擎通过文章中的词语建立倒排索引，并通过统计相同词语的个数等方式来计算查询词与网页的相关性。这种方式只能为用户返回字面上匹配的内容。基于知识增强预训练模型的搜索引擎，通过查询请求和网页内容的统一语义表示，实现了基于语义理解与匹配的搜索，使搜索效果显著提升。

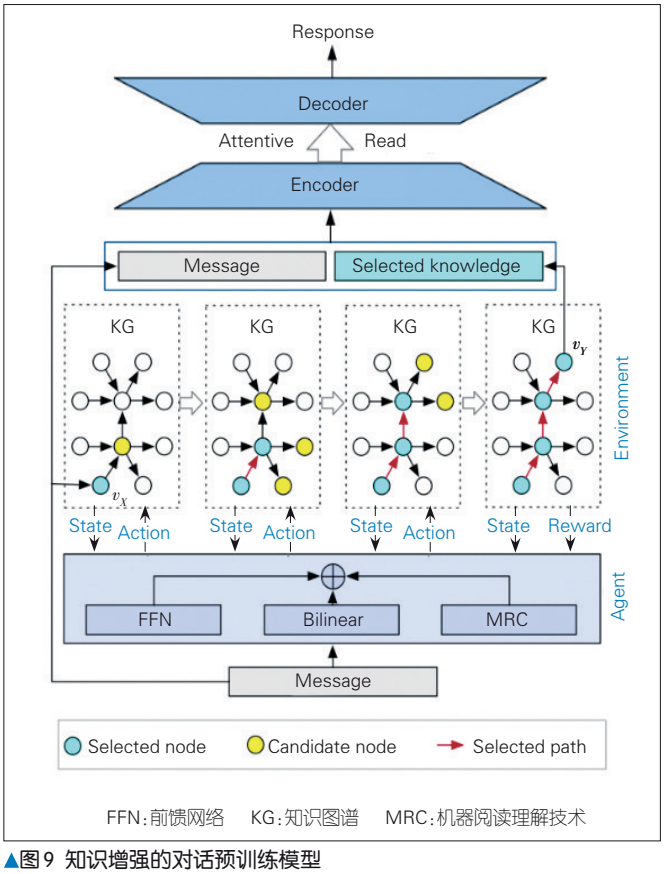
除了应用于搜索引擎的检索阶段和排序阶段之外，文心知识增强模型也能对用户搜索查询的意图进行分析与识别。用户搜索意图识别的准确性将直接影响用户使用搜索引擎的满意度。传统的用户意图识别方法多基于监督学习方法，受限于标注数据的覆盖度，对冷门知识信息搜索查询的识别准确率并不高。而基于文心的用户搜索意图识别方法，能够学

习大量的数据和知识，具备更强的泛化性，使得冷门知识信息搜索意图准确率比传统方法高12%。

3.2 人机对话应用

让机器像人一样有逻辑、有知识、有情感地与人对话，是人机交互的重要发展方向之一。知识增强的对话预训练模型通过对海量无标注数据和大规模知识的学习，使人机对话系统可以更容易模仿人与人的交互方式，让人使用更加自然的方式与机器交流。典型的应用包括智能音箱、智能客服、智能车载等。

文心系列模型包含了基于知识增强的对话预训练模型PLATO[48-50]。基于PLATO模型，我们探索了知识内化和知识外用两种知识增强技术，如图9所示。知识内化是指，在训练阶段，模型将知识信息内化到模型参数中。通过多阶段的模型训练方式来引入大规模通用领域问答知识，可使PLATO融入生成问答能力，进而将问答准确率从3.2%提升至90%。知识外用是指，在推理阶段，模型动态地引入外部知识以指导回复生成。这两种方式能够有效提升PLATO多轮对话的内容丰富度和主题连贯性。



▲图9 知识增强的对话预训练模型

3.3 行业领域应用

知识增强预训练模型在医疗、金融、媒体等人工智能行业中表现出极大的应用价值。

在医疗行业中，中国的医疗卫生事业存在医疗资源不平衡、医生人力短缺等问题。基于知识增强预训练模型构建的临床医疗辅助技术是解决这些问题的关键技术之一。知识增强的医疗语义理解与图推理模型[51-52]，可实现医学知识的计算，并通过患者场景化子图推断，实现可循证的医学决策。该技术突破了以往数据驱动的深度学习技术不可解释的局限，大幅提升了推理决策效果，具备贴合医学临床诊疗思维的优点，改善了临床辅助决策和智能诊前助手等场景应用效果，提高了医护人员临床工作效率。

在金融行业中，知识增强的文心模型被用于金融文本分析，提高了企业对金融信息的处理与决策效率。金融行业需要处理大量的文本信息，例如企业新闻、行业报道、招股书、财报、合同等。在传统模式下，金融从业人员很难从海量文本中获得有效信息。而基于文心模型构建的金融知识计算引擎能够帮助他们从海量的金融文本中快速查找到有用的关键信息。例如，文心模型能够对保险合同中的条款文本进行解析，可实现39个维度的关键信息抽取，使单份合同的处理时间从30 min降低到1 min，能显著提升金融从业人员的工作效率和决策能力。

在媒体行业中，知识增强的文心模型对语言、知识和创作成果进行持续学习，能够实现智能辅助创作。在文章撰写的过程中，基于文心模型的智能创作引擎会对全网热点资讯进行系统分析与计算，为撰稿人提供素材推荐、智能纠错、标题生成、用词润色、文章审校等全方位的帮助。除了自动创作文本外，知识增强的跨模态文心模型实现了以文生图。文心模型可根据文章的文字内容输出具有原创性和艺术性的图片，并将其作为文章的配图使用，进一步丰富内容创作。在知识增强预训练模型的帮助下，智能创作平台将人类从重复劳动中解放出来，有效提升了内容生产的效率和效果。

4 结束语

本文系统阐述了知识增强预训练模型的发展脉络，分析了现有知识增强预训练模型对语言知识、世界知识、领域知识等知识的融合方法，重点介绍了文心知识增强预训练模型的原理、方法和应用效果。通过搜索引擎、人机对话、行业应用3个方面详细介绍了知识增强预训练模型的应用。

知识增强预训练模型已经取得长足发展，但诸多研究方向依然面临巨大挑战。例如，由于知识的稀疏性，现有知识增强预训练模型依旧难以解决逻辑、常识等问题；由于模型

是基于深度神经网络方法来建立的，模型的可解释性、可靠性和可控性仍然较差。因此，如何使模型更具常识性，如何提升模型的可解释性和可靠性，以及如何将跨模态知识、符号化知识与深度学习进行深度融合，都是知识增强预训练模型未来发展的重要方向。

参考文献

[1] SUEN C Y. N-gram statistics for natural language understanding and text processing [J]. IEEE transactions on pattern analysis and machine intelligence, 1979, 1(2): 164-172. DOI: 10.1109/tpami.1979.4766902

[2] BROWM P, PIETRA D, MERCER R. Class-based N-gram models of natural Language [J]. Computational linguistics, 1992, 18(4): 467-480

[3] ODELL J J, VALTCHEV V, WOODLAND P C, et al. A one pass decoder design for large vocabulary recognition [C]//Proceedings of the Workshop on Human Language Technology-HLT '94. Association for Computational Linguistics, 1994: 405-410. DOI: 10.3115/1075812.1075905

[4] MARIÑO J B, BANCHS R E, CREGO J M, et al. N-gram-based machine translation [J]. Computational linguistics, 2006, 32(4): 527-549. DOI: 10.1162/coli.2006.32.4.527

[5] BENGIO Y, DUCHARME D, VINCENT P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3(6): 932-938

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2022-02-15]. https://arxiv. org/abs/1301.3781

[7] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2014: 1532-1543. DOI: 10.3115/v1/d14-1162

[8] 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. DOI: 10.1162/tacl_a_00051

[9] PETERS M, MAUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. [2022-02-15]. https://aclanthology.org/N18-1202.pdf

[10] DEVLIN J, CHANG M, LEE K, et al. Pre-training of deep bidi-rectional transformers for language understanding [EB/OL]. [2022-02-15]. https://aclanthology.org/N19-1423.pdf

[11] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6000-6010

[12] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. [2022-02-15]. https://proceedings. neurips. cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[13] SCHICK T, SCHÜTZE H. Exploiting cloze-questions for few-shot text classification and natural language inference [C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2021: 255-269. DOI: 10.18653/v1/2021.eacl-main.20

[14] SUN Y, WANG S H, FENG S, et al. Ernie 3.0: large-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2107.02137

[15] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration [EB/OL]. [2022-02-15]. https://arxiv. org/abs/1904.09223v1

[16] XIONG W H, DU J F, WANG W Y, et al. Pretrained encyclopedia: weakly supervised knowledge-pretrained language model [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1912.09637

[17] LIU W J, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1909.07606

[18] SUN T, SHAO Y, QIU X, et al. CoLAKE: contextualized language and knowledge embedding [EB/OL]. [2022-02-15]. https://arxiv. org/abs/2010.00309

[19] XIAO D L, LI Y K, ZHANG H, et al. ERNIE-gram: pre-training with explicitly N-gram masked language modeling for natural language understanding [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2010.12148

[20] YE D, LIN Y, DU J, et al. Coreferential reasoning learning for language representation [EB/OL]. [2022-02-15]. https://aclanthology. org/2020.emnlp-main.582.pdf

[21] MILLER G. Wordnet: a lexical database for English [J]. Communications of the ACM, 1995, 38(11):39-41

[22] DONG Z D, DONG Q. HowNet: a hybrid language and knowledge resource [C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2003: 820-824. DOI: 10.1109/NLPKE.2003.1276017

[23] LEVINE Y, LENZ B, DAGAN O, et al. SenseBERT: driving some sense into BERT [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1908.05646

[24] LAUSCHER A, VULIC I, PONTI E, et al. Specializing unsupervised pretraining models for word-level semantic similarity [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1909.02339

[25] JI S X, PAN S R, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications [J]. IEEE transactions on neural networks and learning systems, 2022, 33(2): 494-514. DOI: 10.1109/TNNLS.2021.3070843

[26] WANG X Z, GAO T Y, ZHU Z C, et al. KEPLER: a unified model for knowledge embedding and pre-trained language representation [J]. Transactions of the association for computational linguistics, 2021, 9: 176-194. DOI: 10.1162/tacl_a_00360

[27] ZHANG N, BI Z, LIANG X, et al. Cblue: a Chinese biomedical language understanding [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2106.08087

[28] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2019: 36(4): 1234-1240. DOI: 10.1093/bioinformatics/btz682

[29] HE B, ZHOU D, XIAO J H, et al. BERT-MK: integrating graph contextualized knowledge into pre-trained language models [C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, 2020: 2281-2290. DOI: 10.18653/v1/2020.findings-emnlp.207

[30] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2109.01652

[31] ARIBANDI V, TAY Y, SCHUSTER T, et al. Ext5: Towards extreme multi-task scaling for transfer learning [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2111.10952

[32] SANH V, WEBSON W, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization. [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2110.08207

[33] GU Y, HAN X, LIU Z, et al. PPT: pre-trained prompt tuning for few-shot learning [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2109.04332

[34] SUN Y, WANG S H, LI Y K, et al. ERNIE 2.0: a continual pre-training framework for language understanding [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1907.12412

[35] WANG R Z, TANG D Y, DUAN N, et al. K-adapter: infusing knowledge into pre-trained models with adapters [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2002.01808

[36] WANG S H, SUN Y, XIANG Y, et al. ERNIE 3.0 titan: exploring larger-scale knowledge enhanced pre-training for language understanding and generation [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2112.12731

[37] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1901.02860

[38] DING S Y, SHANG J Y, WANG S H, et al. ERNIE-doc: a retrospective long-document modeling transformer [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2012.15688

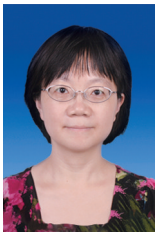
[39] LIU Y, WAN Y, HE L F, et al. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2009.12677

[40] PETERS M E, NEUMANN M, LOGAN R, et al. Knowledge enhanced contextual word representations [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019: 43-54. DOI: 10.18653/v1/d19-1005

[41] XUE L T, CONSTANT N, ROBERTS A, et al. MT5: a massively multilingual pre-trained text-to-text transformer [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2010.11934
[42] ZHANG Z Y, GU Y X, HAN X, et al. CPM-2: large-scale cost-effective pre-trained language models [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2106.10715
[43] OUYANG X, WANG S H, PANG C, et al. ERNIE-M: enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2012.15674
[44] LU J S, BATRA D, PARIKH D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1908.02265
[45] YU F, TANG J J, YIN W C, et al. ERNIE-ViL: knowledge enhanced vision-language representations through scene graph [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2006.16934
[46] LIU Y D, LU W X, CHENG S Q, et al. Pre-trained language model for web-scale retrieval in Baidu search [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 3365-3375. DOI: 10.1145/3447548.3467149
[47] ZOU L X, ZHANG S Q, CAI H Y, et al. Pre-trained language model based ranking in Baidu search [C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. ACM, 2021: 4014-4022. DOI: 10.1145/3447548.3467147
[48] BAO S Q, HE H, WANG F, et al. PLATO: pre-trained dialogue generation model with discrete latent variable [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020: 85-96. DOI: 10.18653/v1/2020.acl-main.9
[49] BAO S, HE H, WANG F, et al. PLATO-XL: exploring the large-scale pre-training of dialogue generation [EB/OL]. [2022-02-15]. https://arxiv.org/abs/2109.09519
[50] LIU Z B, NIU Z Y, WU H, et al. Knowledge aware conversation generation with explainable reasoning over augmented graphs [EB/OL]. [2022-02-15]. https://arxiv.org/abs/1903.10245
[51] YUAN Q, CHEN J, LU C, et al. The graph-based mutual attentive network for automatic diagnosis [C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2020: 3393-3399. DOI: 10.24963/ijcai.2020/469
[52] CHEN J, YUAN Q, LU C, et al. A novel sequence-to-subgraph framework for diagnosis classification [EB/OL]. [2022-02-15]. https://www.ijcai.org/proceedings/2021/0496.pdf

作 者 简 介

王海峰，百度首席技术官、深度学习技术及应用国家工程实验室主任、教授级高工、国际计算机语言学学会（ACL）50余年来首位华人主席、ACL亚太分会创始主席、ACL Fellow、IEEE Fellow、CAAI Fellow、国际欧亚科学院院士；长期从事自然语言处理技术研究及产业化工作；突破知识增强深度语义理解技术，主持研制了产业级深度学习开源开放平台飞桨，实现了智能搜索、机器翻译等大规模产业应用；以第一完成人获国家技术发明奖二等奖1项、国家科技进步奖二等奖1项、中国专利金奖1项等，获光华工程科技奖、全国创新争先奖、首个吴文俊人工智能杰出贡献奖等；发表论文近200篇，获授权专利150余项。



孙宇，百度杰出研发架构师；主要研究领域包括语义理解、语义表示与计算、大规模预训练模型等；曾获国家技术发明奖二等奖1项、中国电子学会科技进步奖一等奖1项、世界人工智能大会SAIL奖等；发表顶级学术会议论文10余篇，获授权专利30余项。

吴华，百度技术委员会主席、教授级高工；主要从事自然语言处理技术研究及产业化工作；曾获国家技术发明奖二等奖1项、国家科技进步奖二等奖1项、中国专利金奖1项，荣获杰出工程师、青年北京学者等奖项；发表论文100余篇，获授权专利100余项。