计算机科学
COMPUTER SCIENCE

基于预训练语言模型的机器翻译最新进展

杨滨瑕, 罗旭东, 孙凯丽

引用本文

杨滨瑕, 罗旭东, 孙凯丽. 基于预训练语言模型的机器翻译最新进展[J]. 计算机科学, 2024, 51(6A): 230700112-8.

YANG Binxia, LUO Xudong, SUN Kaili. Recent Progress on Machine Translation Based on Pre-trained Language Models [J]. Computer Science, 2024, 51(6A): 230700112-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700112

基于预训练语言模型的机器翻译最新进展

杨滨瑕　罗旭东　孙凯丽
广西师范大学计算机科学与工程学院/软件学院　广西 桂林 541004
教育区块链与智能技术教育部重点实验室　广西 桂林 541004
广西多源信息挖掘与安全重点实验室　广西 桂林 541004
(binxiay@stu.gxnu.edu.cn)

摘　要　自然语言处理涉及许多重要主题,其中之一是机器翻译。预训练语言模型,如BERT和GPT,是用于处理包括机器翻译在内的各种自然语言处理任务的先进方法。因此,许多研究人员使用预训练语言模型来解决机器翻译问题。为推动研究向前发展,首先概述了这一领域的最新进展,包括主要的研究问题和基于各种预训练语言模型的解决方案;其次比较了这些解决方案的动机、共性、差异和局限性;然后总结了训练这类机器翻译模型常用的数据集,以及评估这些模型的指标;最后讨论了进一步的研究方向。

关键词:自然语言处理;机器翻译;预训练语言模型;BERT;GPT
中图分类号　TP391

Recent Progress on Machine Translation Based on Pre-trained Language Models

YANG Binxia,LUO Xudong and SUN Kaili
School of Computer Science and Engineering & School of Software,Guangxi Normal University,Guilin,Guangxi 541004,China
Key Laboratory of Blockchain and Intelligent Technology in Education,Ministry of Education,Guilin,Guangxi 541004,China
Guangxi Key Lab of Multi-source Information Mining & Security,Guilin,Guangxi 541004,China

Abstract　Natural language processing(NLP) involves many important topics,one of which is machine translation(MT). Pre-trained language models(PLMs),such as BERT and GPT,are state-of-the-art approaches for various NLP tasks including MT. Therefore,many researchers use PLMs to solve MT problems. To push the research forward,this paper provides an overview of recent advances in this field,including the main research questions and solutions based on various PLMs. We compare the motivations,commonalities,differences and limitations of these solutions,and summarise the datasets commonly used to train such MT models,as well as the metrics used to evaluate them. Finally,further research directions are discussed.

Keywords　Natural language processing,Machine translation,Pre-trained language model,BERT,GPT

1　引言

机器翻译(Machine Translations,MT)在自然语言处理(Natural Language Processing,NLP)中具有重要价值,它使得翻译过程自动化,降低了对人工干预的需求。在过去,机器翻译主要依赖于统计模型,而如今,神经机器翻译(Neural Machine Translation,NMT)得到了更为广泛的应用[1]。神经机器翻译利用带有编码器和解码器的神经网络将源句子转换为目标句子[2]。尽管神经机器翻译改进了统计机器翻译许多的不足之处,如准确性,但它仍面临着一些挑战,例如在处理大数据集时计算需求较高。

目前研究人员正尝试利用预训练语言模型(Pre-trained Language Models,PLMs)来解决现有神经机器翻译存在的问题[3]。通过对大量文本进行无监督训练,预训练语言模型可以获得通用语言表征。由于它们的上下文作为标签,因此预训练语言模型得以使用几乎无限的大规模语料库进行训练。在各种下游任务中,预训练语言模型表现出了优异的性能,这使得它们在机器翻译任务中越来越受欢迎。

本文旨在通过调查最先进的基于预训练语言模型的机器翻译方法,进一步推动对该主题的研究。与其他关于预训练语言模型和机器翻译的调查相比,本文的调查具有独特的优势。例如,Rivera-Trigueros[4]回顾了最常见的机器翻译系统,但提供的基于预训练语言模型的机器翻译信息有限。相比之下,本文的调查为对基于预训练语言模型的机器翻译感兴趣的研究人员提供了有价值的见解。同样,Sun等[3]对预训练语言模型进行了调查,但他们需要关注基于预训练语言模型的机器翻译模型。此外,他们的综述仅包括其综述文章出版前的参考文献,而本文还纳入了一些其出版后的参考文献。Ranathunga等[5]在对低资源神经机器翻译的调查中提到了某些预训练语言模型在机器翻译中的应用。然而,他们的重点是解决低资源神经机器翻译存在的问题,而不是优先考虑预训练语言模型。相反,本文的调查强调了BERT(来自

项目基金:广西多源信息挖掘与安全重点实验室系统性研究课题基金项目(22-A-01-02)
This work was supported by the Guangxi Key Lab of Multi-source Information Mining & Security(22-A-01-02).
通信作者:罗旭东(luoxd@mailbox.gxnu.edu.cn)

变换器的双向编码器表示)[6]、GPT(生成预训练变换器)[7]、XLM(跨语言语言模型)[8]和 BART(双向和自动回归变换器)[9]等预训练语言模型的使用,以及它们在解决不同类型机器翻译问题中的应用。从根本上说,我们对基于预训练语言模型的机器翻译的关注使本文的调查有别于其他专注于不同方面的调查。此外,本文的调查涵盖了在 Ranathunga 等的调查之后发布的机器翻译模型,使其成为一个更新颖的资源。

本文第 2 章回顾了基于 BERT 的机器翻译模型;第 3 章研究了基于 GPT 的机器翻译模型;第 4 章简要概述了基于 BERT&GPT 的机器翻译模型;第 5 章探讨了基于 XLM 的机器翻译模型;第 6 章总结了基于 BART 的机器翻译模型;第 7 章和第 8 章分别总结了常用的数据集和评估指标;第 9 章指出了未来的研究方向;最后总结全文。

2 基于 BERT 的机器翻译

本章将探讨研究人员如何使用 BERT 进行不同类型的机器翻译。表 1 列出了这些方法的引用情况、解决的问题及其优缺点。BERT 是由谷歌开发的一种预训练语言模型,被认为是自然语言处理领域中最著名和最具影响力的模型之一[3]。

表 1 基于 BERT 的机器翻译模型的比较

Table 1 Comparison of BERT-based MT models

参考文献	引用	解决的问题	优点	缺点
Yang 等[10]	120	如何将 BERT 纳入神经机器翻译,因为过多的更新会使 BERT 在训练前忘记已有的知识	BLEU 得分超过了之前的方法	解码部分的结果不佳
Zhu 等[11]	355	如何将 BERT 纳入神经机器翻译	在多种翻译任务上取得最优的结果	推理过程缓慢
Zhang 等[12]	0	利用 BERT 改进神经机器翻译的注意机制,提高神经机器翻译模型的表示能力	能更好地学习句子中的词依赖关系,捕捉句子的内部结构	提取的信息不够精细,推理时间过长
Shavarani 等[13]	8	将 BERT 的语言信息有效地用于神经机器翻译	在保持低计算复杂性的同时提高了翻译质量	不一定适用于低资源数据
Guo 等[14]	8	有效地将 BERT 纳入神经机器翻译任务中	加快了参数效率和解码速度	内存使用过多
Tran[15]	24	在有限的计算预算下,将现有的预训练语言模型从英语转为其他语言	在双语零点任务中的表现优于多语种	使用的 BERT 受语言相似性影响较大
Miyazaki 等[16]	11	解决手语数据短缺的问题,提高开放域内容的翻译质量	解决了数据不足的问题	存在一些无法解决的翻译错误
Üstün 等[17]	27	在多语言无监督机器翻译中有效使用单语数据	允许逐步增加隐形语言	尚未用于解决领域适应问题
Briskilal 等[18]	41	使用预训练语言模型对惯用成语进行翻译	精确度高于单一模型	仅应用于成语翻译,尚未拓展到其他领域
Liu 等[19]	0	使用预训练语言模型和机器翻译检查德国新闻文章的真实性	能有效减少德语翻译成英语时出现的错误	使用粗粒度标签处理
Zhang 等[20]	7	在神经机器翻译任务中最大限度地使用 BERT	在多个翻译任务上表现一流	在推理阶段会产生大量计算成本

2.1 高资源环境

高资源环境指可以获取大量源语言和目标语言的并行数据,并配备其他资源(如计算能力和专业人类注释者)的环境。这些条件有助于开发具有卓越准确性和性能的机器翻译系统,通常使用神经机器翻译模型。在英语、汉语、法语和德语等主要语言之间进行翻译时,通常会遇到高资源环境,因为在这些语言之间可以随时获得大量平行数据。

对于高资源环境下的机器翻译,有几种将 BERT 纳入神经机器翻译系统以提高其性能的方法。这些方法包括将 BERT 与神经机器翻译联合训练[10],将 BERT 与神经机器翻译融合[11],通过整合动态注意力聚合和 BERT 增强神经机器翻译[12],通过 BERT 获取的方面级语义信息增强神经机器翻译[13],以及使用自适应适配器将 BERT 集成到神经机器翻译系统中[14]。实验证明这些方法在各种翻译任务(如英译德、德译英和德译法)中能够提高神经机器翻译系统的翻译质量。

这些方法的共同点是在神经机器翻译背景下整合和使用 BERT,它们探索了将 BERT 纳入神经机器翻译模型的各种方法和技术,以提高翻译质量和性能。这些方法的重点是利用 BERT 的上下文嵌入、动态注意力聚合、语言信息提取和自适应适配器来提高神经机器翻译系统的效率和效果。

虽然上述研究的共同目标是利用 BERT 中编码的知识来增强神经机器翻译模型,但分别提出了不同的方法和技术来实现这一目标。首先,Yang 等于 2020 年提出的 BERT 融合神经机器翻译[10]通过将 BERT 的隐藏状态与编码器和解码器状态融合,将 BERT 集成到神经机器翻译模型中。他们还尝试了不同的融合策略,并提出了两种有效的策略:加法和乘法。其次,Zhu 等于同年提出的 BERT 增强神经机器翻译[11]通过将 BERT 的隐藏状态与源嵌入和目标嵌入串联,将 BERT 集成到神经机器翻译模型中。他们还建议使用微调方法使 BERT 适应特定的翻译任务。同样于 2020 年提出的神经机器翻译中 BERT 的动态注意力聚合机制[12],自适应地将 BERT 的注意力与神经机器翻译模型的注意力结合起来。该方法旨在保留 BERT 自注意机制的优点,同时提高计算效率。2021 年提出的 Linguistic-BERT(L-BERT)[13]从 BERT 中提取语言信息并将其纳入神经机器翻译模型。该方法使用语篇标签、命名实体识别和依赖解析等语言特征来提升翻译性能。2021 年提出的自适应适配器[14]是将 BERT 纳入神经机器翻译的一种有效方法。该方法使用神经机器翻译中的适配器模块,该模块以输入序列为条件,可以根据翻译任务进行调整。

然而,这些方法也表现出一定的局限性。Yang 等[10]没有进行消融研究,也没有进行超参数调整;Zhu 等[11]和 Zhang 等[12]的方法面临推理过程缓慢和需要压缩模型的问题;Shavarani 和 Sarkar[13]提出的方法不适用于低资源语言环境;

Guo 等[14]的方法不适用于多语言或多领域场景。

2.2 低资源环境

在机器翻译领域，低资源环境指用于训练和微调数据驱动机器翻译模型的并行数据量有限的情况。并行数据包括源句和目标句，其中源句为一种语言，目标句为另一种语言的译文。

在低资源环境下，为了解决数据量不足的问题，已经提出了几种方法，包括将 BERT 和 RoBERTa 的预训练语言模型从英语转移到其他语言[15]，在神经机器翻译系统中使用 BERT 作为编码器[16]，引入去噪适配器以增强 mBERT 的交叉注意，从而学习和编码特定语言的信息[17]，使用基于 BERT 和 RoBERTa 的集合模型架构进行中文成语和文本分类[18]，以及将低资源数据集翻译成英语[19]。实验证明，这些方法可以提升机器翻译模型在低资源环境下的性能。

这些方法的共同点如下：1）都使用某种形式的迁移学习或微调，这涉及使用有限的标记数据使预训练语言模型适应新的任务或语言；2）处理机器翻译的特定任务，包括翻译、成语分类和假新闻检测等；3）为不断增长的关于预训练语言模型使用和适应的研究做出了贡献，展示了预训练语言模型在不同场景下的灵活性和适用性；4）证明了微调和迁移学习在使预训练语言模型适应新任务、语言或模式方面的重要性，突出了这些模型的泛化能力。

尽管上述模型都有一个共同的目标，即利用 BERT 中编码的知识来增强低资源环境下的神经机器翻译模型的性能，但它们提出了不同的方法和技术来实现这一目标。2020 年，Tran[15]提出了一种通过在平行数据上微调英语预训练模型来适应其他语言的方法。2020 年，Miyazaki 等[16]使用 BERT 作为编码器探索了从口语到手语的机器翻译。他们的工作是独特的，因为它涉及模态之间的翻译（口语到手语）。2021 年，Üstün 等[17]提出了一种使用去噪适配器的多语言无监督神经机器翻译模型，包括通过微调 mBERT 与去噪自编码器来适应无监督机器翻译。2022 年，Briskilal 和 Subalalitha[18]提出了一个使用 BERT 和 RoBERTa 模型对习语进行翻译的集合模型，以此来应对机器翻译中习语自动检测的挑战。同年，Liu 和 Thoma[19]使用机器翻译将德语文本翻译成英语，然后将假新闻检测技术应用于翻译文本中。

然而，这些方法还存在一定的局限性。Tran 没有考虑各种预训练目标和架构对迁移性能的影响；Miyazaki 等需要充分处理手语的指向特征，这对于表达代词和空间关系至关重要；Üstün 等没有捕捉到不同语言之间的语言多样性；特定领域的局限性限制了 Briskilal 和 Subalalitha 采用的方法；Liu 和 Thoma 需要探索具有更精细标签的分类标准。

2.3 低资源和高资源环境

一些研究人员还研究了如何将预训练语言模型用于低资源语言和高资源语言之间的机器翻译。例如，Zhang 等[20]于 2021 年提出了一种新方法，称为 BERT 联合注意机制（BERT-JAM），可以更好地将强大的 BERT 用于神经机器翻译任务。BERT-JAM 的目的是在神经机器翻译系统的编码器-解码器架构中更好地利用 BERT 丰富的语言知识。这种方法结合了联合关注机制，使神经机器翻译模型的编码器和解码器能够与 BERT 表征进行交互，从而提升翻译性能。此外，该机制有助于更有效地捕捉源语言的语义和句法信息，从而实现准确流畅的翻译。在低资源和高资源环境下的各种翻译任务中，通过实验证实了 BERT-JAM 的一流性能。与此同时，上述方法也存在一定的局限性，例如在推理阶段需要大量的计算成本。

3 基于 GPT 的机器翻译

GPT 是一种基于 Transformer 架构的广为人知且极具影响力的预训练语言模型，由 OpenAI 开发，并且已经在互联网上的大量不同文本数据上进行了预训练。大多数基于预训练语言模型的机器翻译模型使用 BERT，而使用 GPT 的并不多，这可能是因为 GPT 最初并不是为机器翻译开发的。然而，GPT 仍然可以为这项任务进行有效的微调。因此，本章将评述基于 GPT 的机器翻译。表 2 列出了这些方法的引用情况、解决的问题及其优缺点。

表 2 基于 GPT 的机器翻译模型的比较

Table 2 Comparison of GPT-based MT models

参考文献	引用	解决的问题	优点	缺点
Han 等[21]	15	使用生成性语言模型进行无监督翻译	能实现高效的无监督翻译	仅适用于英语和法语之间的翻译
Tan 等[22]	22	减少预训练和翻译之间的差异	显著提升预训练语言模型的翻译性能	计算成本高

下面介绍将 GPT 用于机器翻译的一般方法：

1）数据准备：从平行语料库开始，该语料库包含两种语言的对译句子，如一个英语句子和它的法语译文。

2）预处理：对源语和目标语句子进行标记化处理。根据语言的特殊性，可能还需要应用其他预处理步骤，如小写、去除标点符号等。

3）训练设置：对 GPT 模型的输入进行设置，使其能够学习翻译任务。通常情况下，这包括将源句和目标句连接起来，中间加上一个特殊标记。对模型进行训练，使其能够根据前面的所有单词预测下一个单词。

4）训练：使用 Transformer 架构的变体对模型进行微调。在训练过程中，GPT 学会关联源句和目标句，从而有效地学会从源语翻译为目标语。

5）翻译：要翻译一个新句子，需要将源句子输入模型中，然后让它生成一个续句，直到输出一个停止标记，标志着翻译的结束。

这只是对翻译过程的简单描述，其中可能涉及许多复杂因素，尤其是在处理多种语言、不同句子长度以及 Transformer 架构的限制（如最大序列长度）时。

最近提出的使用 GPT 的两种机器翻译模型如下：第一个模型是 Han 等[21]在 2021 年提出的，旨在使用生成式预训练语言模型（如 GPT）执行无监督神经机器翻译，而不依赖于标记的平行数据进行训练。该模型在翻译任务中使用去噪自动编码器和反向翻译组合对预训练语言模型进行微调，并利用翻译提示候选、迭代反向翻译和模型集合来提高无监督翻译的质量。在 WMT14 英译法基准测试中的实验取得了较高的 BLEU 分数，这表明使用 GPT 的无监督神经机器翻译是开发神经机器翻译系统的一个有前途的方向，而无须依赖标记的平行数据。第二个模型是 Tan 等[22]在 2022 年提出的多阶段提示（MSP），它将翻译过程分为几个阶段，并在每个阶段独立应用不同的顺序提示，以提升预先训练的多语言 GPT

(mGPT)模型在翻译任务中的性能。在不同语言对上的实验表明,预训练语言模型的性能有了显著提高。

这两个模型的共同点在于:1)都旨在提升 GPT 在翻译任务中的性能,以展示 GPT 在机器翻译领域的潜力;2)都使用某种形式的微调或适应 GPT,以改善其在翻译任务中的表现;3)都为不断增长的有关在机器翻译中使用和调整预训练语言模型的研究做出了贡献,展示了这些模型在各种翻译场景中的灵活性和适用性。

这两种实现机器翻译的模式都是基于 GPT 的,但它们在某种程度上是不同的。第一种模式是对模型进行微调,使用多阶段的提示来指导翻译过程,解决了单一提示方法存在局限性的问题。第二种模式是使用 GPT 的无监督神经机器翻译模型,该方法不依赖于并行数据进行训练,而是侧重于利用预先训练的生成模型中存储的知识来执行无监督翻译。

同时,这些方法也存在一定的局限性。Tan 等未能考虑不同预训练语言模型和线索类型对翻译性能的影响,Han 等的方法无法有效处理低资源语言对。一般来说,尽管 GPT 可用于机器翻译,但像谷歌的 Transformer 这样的专用翻译模型(使用专为翻译设计的编码器-解码器结构)在这一任务中的表现往往优于 GPT。

4 基于 BERT&GPT 的机器翻译

本章将简要介绍基于 BERT 和 GPT 的机器翻译模型。表 3 列出了这些方法的引用情况、解决的问题和优缺点。

表 3 基于 BERT 和 GPT 的机器翻译模型的比较

Table 3 Comparison of BERT&GPT-based MT models

参考文献	引用	解决的问题	优点	缺点
Weng 等[23]	57	将从预训练语言模型中获取的知识应用于神经机器翻译,使预训练语言模型的上下文知识很好地与神经机器翻译结合	有效利用预训练知识提高翻译质量	动态融合过程生成的特定任务表示不完整且包含噪声
Zhang 等[24]	19	从网络抓取语料中过滤出噪声句对,用于训练机器翻译模型	在中文翻译数据集上取得最优的结果	尚未能应用于低资源数据
Sawai 等[25]	4	使用 GPT-2 的句子生成器为神经机器翻译生成额外的数据,该生成器生成的句子与原始句子的特征相似	性能随数据量的增加而提高	GPT-2 的句子生成需要很长时间,准备足够的数据是一个障碍

近年来,许多学者建议将 BERT 和 GPT 整合到神经机器翻译模型中以提升其性能。2020 年,Weng 等[23]提出使用预先训练的模型,如 BERT 和 GPT,从一般语言数据中获取知识,然后将其转移到神经机器翻译网络中。该方法包括两个模块:1)动态融合机制,将一般知识中的特定任务特征融合到神经机器翻译中;2)知识提炼范式,在神经机器翻译训练过程中不断学习语言知识。同年,Zhang 等[24]提出了一种利用预训练语言模型从网络抓取语料中过滤噪声句对的模型。该模型使用 BERT 的多语言能力测量句子并行性,并使用 GPT 作为域过滤器,在日中并行语料库中的表现明显优于基线。2021 年,Sawai 等[25]提出了一种使用 GPT-2 进行句子扩充以增强神经机器翻译模型性能的新方法。他们使用 GPT-2 生成现有句子的转述,并将其纳入训练过程,从而提高了神经机器翻译模型的泛化能力和翻译质量。他们的模型优于传统模型,并且性能随着句子数量的增加而提升。

这 3 种模型的共同点如下:1)都旨在通过结合预训练语言模型来提高神经机器翻译模型的性能,展示了这些模型在机器翻译领域的有效性;2)都采用了将预训练语言模型适应或集成到神经机器翻译模型中的形式,展示了这些模型在各种翻译场景中的灵活性和适用性;3)都进行了大量的实验来证明其有效性,展示了预训练语言模型在提高神经机器翻译模型性能方面的潜力。

这 3 种模型的区别如下:1)各自使用的具体方法和技术有很大不同。Weng 等侧重于适配器和知识提取,Zhang 等侧重于并行语料库过滤,Sawai 等侧重于句子增强。2)从不同方面探讨了预训练语言模型在神经机器翻译任务中的应用,例如使用预训练语言模型的语言知识、过滤平行数据及使用转述扩展训练数据集。

然而,这些方法也存在一定的局限性。Weng 等仅提供了与其他在神经机器翻译中使用预训练语言模型的先进模型的部分比较,这使得全面评估他们的模型的优势具有一定的挑战性;Zhang 等在单个语言对(日语-汉语)上评估了他们的模型,这限制了他们的发现在其他语言对上的通用性;Sawai 等的方法在生成句子时耗时较长,这为准备足够数据带来了困难。

5 基于 XLM 的机器翻译

本章将简要介绍基于预训练语言模型 XLM 的机器翻译模型,表 4 列出了这些方法的引用情况、解决的问题及其优缺点。XLM(跨语言模型)是一种预训练语言模型,由 Facebook 人工智能研究院开发,旨在促进自然语言处理任务中的跨语言理解和迁移学习。XLM 在大量多语言文本数据上进行训练,使其能够学习单词和句子的表征,从而捕捉不同语言之间的语言相似性和可迁移知识。

表 4 基于 XLM 的机器翻译模型的比较

Table 4 Comparison of XLM-based MT models

参考文献	引用	解决的问题	优点	缺点
Rubino 和 Sumita[26]	15	使用预先训练的句子编码器改进机器翻译质量评估,解决机器翻译质量评估中注释数据集较少的问题	不依赖注释数据	无法定位被误译的标记词
Li 等[27]	15	为 WMT20 新闻翻译任务建立监督和非监督神经机器翻译系统	有效增强机器翻译系统低资源翻译的能力	使用更多数据时效果无明显改善
Chen 等[28]	15	使用多语言预训练编码器实现神经机器翻译的零次跨语言转移	有效提升多语言神经机器翻译性能	不适用于单语言数据
Ma 等[29]	44	如何通过增强预先训练好的多语言编码器来预先训练编码器-解码器模型,用于语言生成和翻译	性能优于其他多语言预训练语言模型和神经机器翻译模型	模型规模有待扩大
Sun 等[30]	9	利用跨语言预训练语言模型开发统一的多语言语法纠错策略	实现通用的多语言语法纠错	受到翻译语料规模的数量限制,无法生成无限数量的纠错句对

研究人员已经将预训练语言模型 XLM 编码器用于各种自然语言处理任务中，包括机器翻译任务。2020 年，Rubino 和 Sumita[26]专注于使用中间自监督学习的方法来提高句子和单词级别的机器翻译质量估计。他们的模型改变了跨语言模型(XLM)的翻译语言模型训练目标，使预训练模型面向目标，该方法不依赖于注释数据。同年，Li 等[27]针对 WMT20 新闻翻译任务提出了有监督和无监督的神经机器翻译系统。与第一个方法不同的是，Li 等更侧重于具体翻译任务中的实际实现和性能。2021 年，Chen 等[28]提出了一种跨语言神经机器翻译模型 Sixt，其使用 XLM-R 初始化编码器和解码器嵌入。该模型设计用于多语种预训练编码器的神经机器翻译零次跨语言传输。与前两种方法不同的是，Chen 等强调零次跨语言迁移学习。同年，Ma 等[29]提出了一种名为 DeltaLM 的新型多语言预训练语言模型。它使用现有的预训练语言模型作为编码器，并增加了解码器作为编码器的附加层。这项工作的突出之处在于它采用了独特的方法，通过增强预先训练的多语言编码器来进行语言生成和翻译的预训练。2022 年，Sun 等[30]着重于应用预训练语言模型 InfoXLM 来纠正非英语语言的语法，以提高机器翻译的准确性。该模型由 InfoXLM 初始化，并使用其粗略翻译作为初始数据。该方法与其他方法的不同之处在于使用预先训练好的跨语言语言模型进行语法纠错。

这 5 种模型的共同点如下：1)都采用了预训练语言模型，如 XLM，XLM-R 和 InfoXLM，以提高机器翻译的质量；2)都讨论了神经机器翻译系统的各个方面和改进，无论是使用自监督学习，引入新的神经机器翻译技术，应用零点跨语言转移，还是增强现有的神经机器翻译模型；3) 每项研究都直接或间接地涉及多语言和跨语言迁移学习，提高了不同语言对之间翻译的准确性；4)都旨在推动机器翻译领域的发展，无论是通过改进翻译质量评估、完善语法校正，还是提高多语言预训练编码器的效率。

同时，这些方法也存在一定的局限性。例如，Rubino 和 Sumita 提出的方法计算成本较高；Li 等未对无监督机器翻译模型进行详细分析；Chen 等的方法仅在零样本跨语言转移任务中进行了评估，且使用的是从英语到其他语言的并行数据；Ma 等的模型规模需扩大以涵盖其他自然语言理解任务；Sun 等的方法则受到可用翻译语料库规模的限制。

6 基于 BART 的机器翻译

本章将简要介绍基于 BART 的机器翻译模型，表 5 列出了这些方法的引用情况、解决的问题及其优缺点。预训练语言模型 BART 是 Facebook AI Research 专门为多语言应用而设计的。

表 5 基于 BART 的机器翻译模型的比较

Table 5 Comparison of BART-based MT models

参考文献	引用	解决的问题	优点	缺点
Liu 等[31]	1027	提高神经机器翻译在多语言和低资源场景下的性能	在句子和文档层面有效改进了监督和非监督 机器翻译	部署成本高
Wang 等[32]	10	如何在 8 个语言方向上提高生物医学文本的翻译质量：英语-德语、英语-法语、英语-西班牙语和英语-俄语	有效提高翻译质量	未使用大规模的域内数据训练
Dabre 等[33]	29	提高印度语自然语言生成(NLG)的性能	在翻译资源极少的情况下仍能保持良好性能	不适用于长文本
Rippeth 等[34]	2	控制将英语口语翻译成 6 种具有不同语法形式标记的语言的输出形式：德语、法语、西班牙语、俄语、印地语和泰米尔语	实现的翻译质量和对翻译结果格式的控制接近专用翻译模型	翻译质量落后于大型双语模型

近年来，基于 BART 的多语种机器翻译取得了一些进展。2020 年，Liu 等[31]将 BART 扩展为 mBART，这是一种用于神经机器翻译的多语言去噪预训练方法，可以针对任何语言对进行微调，而无须针对特定任务或特定语言进行修改。2021 年，Wang 等[32]使用 mBART 将预训练和微调范式应用于生物医学翻译任务。2022 年，Dabre 等[33]介绍了 IndicBART，这是一种用于印度语自然语言生成的预训练语言模型，重点关注 11 种印度语和英语的翻译任务。同年，Rippeth 等[34]基于 mT5 和 mBART 研究了如何利用最少的资源控制多种语言机器翻译的语法形式，使用加法向量干预来编码风格。实验结果表明，这些方法可以提高机器翻译和语法控制的性能。

这些方法的共同点如下：1)都探讨了如何使用预训练语言模型技术来增强机器翻译系统；2)都针对机器翻译的一个特定方面或挑战进行了研究，例如多语种去噪预训练、印度语机器翻译、生物医学翻译和翻译形式控制；3)各自都进行了实验和评估，展示了其有效性和性能。

这些方法的不同之处在于它们在机器翻译领域内的具体重点和目标。具体来说，Liu 等专注于去噪预训练技术，以增强神经机器翻译模型，特别是在多语言场景中；Dabre 等介绍了 IndicBART，一种专门用于生成印度语自然语言文本的预训练模型；Wang 等展示了他们的机器翻译系统在生物医学领域的性能；Rippeth 等探讨了如何使用多语言预训练语言模型来控制翻译文本的形式，提供了生成不同形式翻译的方法。虽然这些方法都与机器翻译和语言处理有关，但它们在具体目标上有所不同，例如去噪预训练、专注于印度语言、解决生物医学翻译问题及控制翻译形式。

此外，这些方法还存在一定的局限性。Liu 等的模型在

1) https://statmt.org/wmt20/results.html

2) https://www.statmt.org/europarl/

3) https://conferences.unite.un.org/UNCorpus/

4) http://www.opensubtitles.org

5) https://iwslt.org/

6) https://paracrawl.eu/

7) https://commoncrawl.org/

8) https://gitcode.net/mirrors/tensorflow/tensor2tensor?utm_source=csdn_github_accelerator

9) https://www.euromatrixplus.net/multi-un/

生产中的部署成本较高；Dabre 等的方法对计算要求较高，不能有效处理长文档，且他们关注的印度语言可能无法直接适用于其他语系或语对；Wang 等的方法受到特定领域的限制；Rippeth 等的方法容易受到潜在混淆角色的影响。

7 常用数据集

由于神经机器翻译研究人员或从业人员可能需要使用不同的数据集进行训练和评估，因此本章列出了一些常用的数据集。

1）WMT[1)][35]：这是每年一次的机器翻译评估活动，其中包含不同语言对的各种数据集。其中，新闻评论数据集是最常用的。

2）Europarl[2)][36]：这是一个来自欧洲议会记录的平行语料库，有 21 种欧洲语言版本。

3）UNPC[3)][37]：该语料库包含联合国官方文件，是训练神经机器翻译模型的重要来源，包括 6 种语言——阿拉伯语、汉语、英语、法语、俄语和西班牙语。

4）OpenSubtitles[4)][38]：这个数据集包含电影和电视字幕，适用于训练神经机器翻译系统处理非正式语言。该数据集涵盖多种语言，且具有非正式性质。

5）IWSLT[5)]：这些数据集是 TED 演讲的转录和翻译，有多种语言版本。

6）ParaCrawl[6)][39]：该项目对网络进行抓取，并生成 23 种欧洲语言的平行语料库。

7）Common Crawl[7)]：这是另一个网络抓取项目，可用于构建平行语料库。

8）T2T[8)][40]：Tensor2Tensor（T2T）提供了多个翻译数据集，包括 WMT English-to-German 和 WMT English-to-French 等。

9）MultiUN[9)][41]：从联合国文件中创建的多语言平行语料库。

在选择用于训练或评估的数据集之前，研究人员和开发人员必须考虑任务的具体情况以及所使用的语言。有些数据集可能比其他数据集更适合他们的需求。

8 评估指标

评估神经机器翻译模型的性能时，通常会采用多种指标。每种指标都有其优缺点，选择指标可能取决于具体应用或研究问题。以下是一些常用的指标。

1）BLEU（BiLingual Evaluation Understudy）：BLEU 是神经机器翻译中最广泛使用的度量标准。它通过测量翻译句子和参考句子之间 n-grams 的重叠度来评估模型性能。BLEU 得分范围从 0～1，1 为最佳。

2）NIST（National Institute of Standards and Technology）：NIST 是 BLEU 度量的扩展，它为频率较低的词序列赋予额外权重。与 BLEU 一样，NIST 得分越高越好。

3）METEOR（Metric for Evaluation of Translation with Explicit ORdering）：METEOR 是另一个扩展 BLEU 的指标。除了精确度和召回率，METEOR 还考虑了同义词、词干和意译匹配，因此其比 BLEU 和 NIST 更复杂。

4）TER（Translation Edit Rate）：TER 衡量将系统输出修改为一个参考文献所需的编辑次数。TER 分数越低越好。

5）ROUGE（Recall-Oriented Understudy for Gisting Evaluation）：ROUGE 最初用于评估自动摘要，也可用于神经机器翻译。ROUGE-N 指 n-grams 的重叠，而 ROUGE-L 考虑的是最长的共同子序列。

6）WER（Word Error Rate）：WER 是将假设变为参考所需的最小插入、删除和替换数与参考中的字数之比。

7）ChrF（Character n-gram F-score）：ChrF 在字符级别上比较系统输出和参考文献，这对于单词分割复杂或模糊的语言很有帮助。

8）BLEURT（Bilingual Evaluation Understudy with Representations from Transformers）：BLEURT 是一种学习度量，它使用 BERT 嵌入，并在一个包含人类翻译排名的数据集上进行训练。

9）BERTScore：BERTScore 将两个句子的相似度计算为其单词嵌入之间的余弦相似度之和。

10）COMET（Cross-lingual Optimised Metric for Evaluation of Translation）：COMET 旨在更好地与人类对翻译质量的判断相关联。

通常情况下，没有一种度量标准是完美的，人工评估通常被认为是黄金标准。在实际应用中，这些指标通常会与人工评估相结合。传统上，BLEU 和 METEOR 等指标依赖于翻译文本与参考译文之间的 n-grams 匹配。然而，它们可能无法充分捕捉翻译质量的更复杂方面，如意义保持和流畅性。基于此，COMET 应运而生，旨在解决这些局限性。

9 未来挑战

通过对基于预训练语言模型的机器翻译文献的分析，我们认为未来的主要挑战如下：

1）在基于预训练语言模型的机器翻译这一主题上，大多数研究者在类似的语言对上验证他们的方法，例如英译法和英译德，很少考虑语言的多样性。虽然一些研究者，如 Dabre 等[33]已经开始朝这个方向努力，但在低资源语言对之间基于预训练语言模型的机器翻译方面，还需要更多的努力。

2）尝试使用更多的预训练语言模型，特别是更新的预训练语言模型和非英语的预训练语言模型进机器翻译。例如，有许多中文预训练语言模型[3]，但基于这些中文预训练语言模型的机器翻译模型却很少。在现有的预训练语言模型中，哪些预训练语言模型最适合哪些应用领域或任务，也是值得研究的。

3）在传统的机器翻译过程中，文本翻译过程和源信息（如图像）是相互独立的，其翻译结果可能丢失源中包含的重要信息。Kwon 等[42]尝试结合卷积神经网络来解决这一问题，但还需要更多的努力。因此，将多模态特征集成到机器翻译的预训练语言模型中，实现多模态神经机器翻译，以提高翻译质量，获得更智能的翻译，是非常值得的。

4）基于预训练语言模型的先进机器翻译系统翻译质量总体较好，但有时仍会出现一些简单的错误。这是因为机器翻译系统获取翻译知识，特别是常识的能力不足。为了解决这个问题，Sun 等[30]试图获取良好的训练数据。然而，从 20 世纪 90 年代的统计机器翻译到今天的神经机器翻译，从数据中学习特征表征已经达到了最大化。因此，利用外部知识增强预训练语言模型的机器翻译能力是未来提高基于预训练语

言模型的机器翻译性能的关键。

结束语 预训练语言模型作为自然语言处理的核心之一,近年来在各种自然语言处理任务,尤其是机器翻译中取得了巨大的成功。为了促进基于预训练语言模型的机器翻译的进一步研究,本文调查了各种最先进的预训练语言模型在机器翻译中的应用。研究人员运用 BERT,GPT,XLM 和 BART 等模型,不仅解决了传统机器翻译中存在的问题,还对这些方法进行了优化与改进。这些方法不仅适用于高资源语言环境,也能够在低资源环境和跨语言翻译任务中发挥作用,甚至能够应用于手语机器翻译领域。大多数机器翻译任务使用大型数据集,如英译德和英译法(高资源语言)。对于这些方法,本文讨论了动机(即作者想要解决什么问题)以及如何使用各种预训练语言模型来解决这些问题。对于本文讨论的方法,我们比较了它们的共性、差异和局限性,以便未来的研究人员能够为他们的项目选择最合适的方法或开发自己的方法。此外,本文还总结了常用的数据集和评估指标,并确定了未来研究的 4 个方向。其中,利用外部知识增强预训练语言模型的机器翻译功能最为重要。

参 考 文 献

[1] BAHDANAU D,CHO K H,BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// 3rd International Conference on Learning Representations(ICLR). 2015.

[2] ZHANG Z,WU S,JIANG D,et al. BERT-JAM:Maximizing the utilization of BERT for neural machine translation [J]. Neurocomputing,2021,460:84-94.

[3] SUN K,LUO X,LUO M Y. A survey of pretrained language models[C]// International Conference on Knowledge Science,Engineering and Management. Cham: Springer International Publishing,2022:442-456.

[4] RIVERA-TRIGUEROS I. Machine translation systems and quality assessment: a systematic review [J]. Language Resources and Evaluation,2022,56(2):593-619.

[5] RANATHUNGA S,LEE E S A,SKENDULI M P,et al. Neural machine translation for low-resource languages:A survey [J]. ACM Computing Surveys,2023,55(11):1-37.

[6] KENTON,DEVLIN J,CHANG M W,et al. BERT:Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT. 2019:4171-4186.

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

[8] CONNEAU A,LAMPLE G. Cross-lingual language model pretraining[C]// Advances in Neural Information Processing Systems. 2019,32.

[9] LEWIS M,LIU Y,GOYAL N,et al. BART:Denoising Sequence-to-Sequence Pre-training for Natural Language Generation,Translation,and Comprehension[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:7871-7880.

[10] YANG J,WANG M,ZHOU H,et al. Towards making the most of BERT in neural machine translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:9378-9385.

[11] ZHU J,XIA Y,WU L,et al. Incorporating bert into neural machine translation[C]// 8rd International Conference on Learning Representations(ICLR 2020). 2020.

[12] ZHANG J R,LI H Z,SHI S M,et al. Dynamic Attention Aggregation with BERT for Neural Machine Translation[C]// 2020 International Joint Conference on Neural Networks(IJCNN). IEEE,2020:1-8.

[13] SHAVARANI H S,SARKAR A. Better Neural Machine Translation by Extracting Linguistic Information from BERT[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. 2021:2772-2783.

[14] GUO J,ZHANG Z,XU L,et al. Adaptive adapters:An efficient way to incorporate BERT into neural machine translation [J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2021,29:1740-1751.

[15] TRAN K. From English to foreign languages:Transferring pre-trained language models [EB/OL]. https://arxiv. org/abs/2002. 07306.

[16] MIYAZAKI T,MORITA Y,SANO M. Machine translation from spoken language to Sign language using pre-trained language model as encoder[C]// Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages:Sign Language Resources in the Service of the Language Community,Technological Challenges and Application Perspectives. 2020:139-144.

[17] ÜSTÜN A,BÉRARD A,BESACIER L,et al. Multilingual Unsupervised Neural Machine Translation with Denoising Adapters [C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021:6650-6662.

[18] BRISKILAL J,SUBALALITHA C N. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa [J]. Information Processing & Management, 2022, 59(1): 102756.

[19] LIU J,THOMA S. GERMAN TO ENGLISH:Fake News Detection with Machine Translation [J]. Lecture Notes in Informatics,Gesellschaft für Informatik,2022,3457:1-8.

[20] ZHANG Z,WU S,JIANG D,et al. BERT-JAM:Maximizing the utilization of BERT for neural machine translation [J]. Neurocomputing,2021,460:84-94.

[21] HAN J M,BABUSCHKIN I,EDWARDS H,et al. Unsupervised neural machine translation with generative language models only [EB/OL]. https://arxiv. org/abs/2110. 05448.

[22] TAN Z,ZHANG X,WANG S,et al. MSP:Multi-Stage Prompting for Making Pre-trained Language Models Better Translators[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(Volume 1:Long Papers). 2022:6131-6142.

[23] WENG R,YU H,HUANG S,et al. Acquiring knowledge from pre-trained model to neural machine translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:9266-9273.

[24] ZHANG B,NAGESH A,KNIGHT K. Parallel Corpus Filtering via Pre-trained Language Models[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:8545-8554.

[25] SAWAI R,PAIK I,KUWANA A. Sentence Augmentation for

Language Translation Using GPT-2 [J]. Electronics, 2021, 10(24):3082.
[26] RUBINO R, SUMITA E. Intermediate self-supervised learning for machine translation quality estimation[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020:4355-4360.
[27] LI Z, ZHAO H, WANG R, et al. SJTU-NICT's Supervised and Unsupervised Neural Machine Translation Systems for the WMT20 News Translation Task[C]// Proceedings of the Fifth Conference on Machine Translation. 2020:218-229.
[28] CHEN G, MA S, CHEN Y, et al. Zero-Shot Cross-Lingual Transfer of Neural Machine Translation with Multilingual Pretrained Encoders[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 15-26.
[29] MA S, DONG L, HUANG S, et al. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders[EB/OL]. https://arxiv.org/abs/2106.13736.
[30] SUN X, GE T, MA S, et al. A unified strategy for multilingual grammatical error correction with pre-trained cross-lingual language mode[EB/OL]. https://arxiv.org/abs/2201.10707.
[31] LIU Y, GU J, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation [J]. Transactions of the Association for Computational Linguistics, 2020, 8:726-742.
[32] WANG X, TU Z, SHI S. Tencent AI lab machine translation systems for the WMT21 biomedical translation task[C]// Proceedings of the Sixth Conference on Machine Translation. 2021: 874-878.
[33] DABRE R, SHROTRIYA H, KUNCHUKUTTAN A, et al. IndicBART: A Pre-trained Model for Indic Natural Language Generation[C]// Findings of the Association for Computational Linguistics: ACL 2022. 2022:1849-1863.
[34] RIPPETH E, AGRAWAL S, CARPUAT M. Controlling Translation Formality Using Pre-trained Multilingual Language Models[C]// Proceedings of the 19th International Conference on Spoken Language Translation(IWSLT 2022). 2022:327-340.
[35] LOIC B, MAGDALENA B, ONDREJ B, et al. Findings of the 2020 conference on machine translation[C]// Proceedings of the Fifth Conference on Machine Translation. 2020:1-55.
[36] KOEHN P. Europarl: A parallel corpus for statistical machine translation[C]// Proceedings of machine translation summit x: papers. 2005:79-86.
[37] ZIEMSKI M, JUNCZYS-DOWMUNT M, POULIQUEN B. The united nations parallel corpus v1.0[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation(LREC'16). 2016:3530-3534.
[38] LISON P, TIEDEMANN J. Open Subtitles 2016: extracting large parallel corpora from movie and TV subtitles[C]// 10th Conference on International Language Resources and Evaluation (LREC'16). European Language Resources Association, 2016: 923-929.
[39] BAÑÓN M, CHEN P, HADDOW B, et al. ParaCrawl: Web-scale acquisition of parallel corpora[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:4555-4567.
[40] VASWANI A, BENGIO S, BREVDO E, et al. Tensor2Tensor for Neural Machine Translation[C]// Proceedings of the 13th Conference of the Association for Machine Translation in the Americas(Volume 1: Research Track). 2018:193-199.
[41] EISELE A, CHEN Y. MultiUN: A multilingual corpus from united nation documents[C]// LREC. 2010.
[42] KWON S, GO B H, LEE J H. A text-based visual context modulation neural model for multimodal machine translation [J]. Pattern Recognition Letters, 2020, 136:212-218.

YANG Binxia, born in 1999, postgraduate. Her main research interest is natural language processing(NLP).

LUO Xudong, born in 1963, Ph.D, distinguished professor, Ph.D supervisor. His main research interests include natural language processing, intelligent decision-making, game theory, automated negotiation and fuzzy logic.