

文章编号: 1003-0077(2024)06-0001-23

预训练神经机器翻译研究进展分析

曹智泉, 穆永誉, 肖 桐, 李 北, 张春良, 朱靖波

(东北大学 自然语言处理实验室, 辽宁 沈阳 110819)

摘 要: 神经机器翻译(NMT)模型通常使用双语数据进行监督训练,而构建大规模双语数据集是一个巨大挑战。相比之下,大部分语言的单语数据集较为容易获取。近年来,预训练模型(PTM)能够在海量的单语数据上进行训练,从而得到通用表示知识,来帮助下游任务取得显著的性能提升。目前基于预训练的神经机器翻译(PTNMT)在受限资源数据集上已被广泛验证,但如何高效地在高资源 NMT 模型中利用 PTM 仍亟待研究。该文致力于对 PTNMT 的现状和相关问题进行系统性的整理和分析,从引入 PTM 的预训练方法、使用策略以及特定任务等角度对 PTNMT 方法进行详细的分类,并对 PTNMT 方法解决的问题进行总结,最后对 PTNMT 的研究进行展望。

关键词: 自然语言处理;预训练模型;神经机器翻译

中图分类号: TP391

文献标识码: A

Pre-trained Neural Machine Translation: Progress and Analysis

CAO Zhiqun, MU Yongyu, XIAO Tong, LI Bei, ZHANG Chunliang, ZHU Jingbo

(NLP Laboratory, Northeastern University, Shenyang, Liaoning 110819, China)

Abstract: Neural machine translation (NMT) models are usually trained using bilingual data. Building large-scale bilingual datasets is a huge challenge. In contrast, large-scale monolingual datasets for most languages are easier to construct. Pre-trained models (PTM) proposed in recent years can be trained on massive monolingual data. The generic representation of knowledge learned through pre-training helps achieve significant performance gains in downstream tasks. Currently pre-trained neural machine translation (PTNMT) has been extensively validated on resource-constrained datasets, but how to efficiently utilize PTM on high-resource NMT remains to be discussed. This paper focuses on reviewing and analyzing the current state and related problems of PTNMT, classifying PTNMT methods in terms of PTM's pre-trained methods, strategies, or specific tasks. We summarize the problems solved by PTNMT's methods, and conclude with a future outlook on PTNMT research.

Keywords: natural language processing; pre-trained model; neural machine translation

0 引言

机器翻译是自然语言处理领域中一个重要的研究方向,其主要实现的功能是将源语言翻译成目标语言。近年来,随着深度学习技术在人工智能领域的发展,神经机器翻译(Neural Machine Translation, NMT)已经成为机器翻译方向中的主流方法。从

2013 年发展至今,研究人员提出许多模型来实现 NMT,例如,CNN^[1]、RNN^[2]、Transformer^[3]等,最终基于自注意力网络设计的 Transformer 成为了 NMT 的主流范式^[4]。

Transformer 的出现,不仅带动了 NMT 的发展,也带动了整个自然语言处理领域的发展。其中,受 Transformer 影响出现的大规模预训练模型(Pre-Trained Model, PTM)在许多自然语言处理任

收稿日期: 2022-12-05 定稿日期: 2023-05-10

基金项目: 国家自然科学基金(62276056);科技部科技创新 2030—“新一代人工智能”重大项目(2020AAA0107904);云南省科技厅科技计划项目(202103AA080015);中央高校基本科研业务费项目(N2216016, N2216001, N2216002);111 引智基地(B16009)

务上都取得了先进的性能。大量工作表明^[5-9],在大规模无标签数据上训练的 PTM 可以学习到“通用表示知识”,这使 PTM 拥有强大的语言理解和生成能力。PTM 的“通用表示知识”也引起了机器翻译研究人员的关注,研究人员期望通过引入 PTM 来提高 NMT 的性能。这种将 PTM 引入 NMT 的方法被称为预训练神经机器翻译(Pre-trained Neural Machine Translation, PTNMT)。

与不引入 PTM 的 NMT 相比,PTNMT 的优势可以总结为以下几点:

- (1) PTM 可以使 NMT 模型很好地利用无标签数据;
- (2) PTM 可以提高 NMT 模型的泛化能力和鲁棒性;
- (3) PTM 可以帮助 NMT 搭建通用的翻译模型。

当前,PTNMT 存在大量的工作,但尚不存在独立地对其进行总结的综述文章,已有的 PTNMT 相关整理工作都是以一节的形式出现在 PTM 综述文章里。一部分工作^[10-11]仅简单地对 PTNMT 进行介绍,另一部分工作^[12]则是从单语言和多语言角度对 PTNMT 的部分工作进行分类说明。虽然也存在对 PTNMT 整体发展进行描述的报告^[13],但这些工作都缺少对 PTNMT 相关工作的系统性整理和分析。本文主要以 PTNMT 的构建、问题分析和应用为切入点,对相关工作进行详细的分类,侧重于对不同方法进行对比,涵盖了最新的相关工作,分析了 PTNMT 的优缺点及其对其他任务的启发,首次完成对 PTNMT 相关工作的系统性整理和分析工作。

本文组织结构如下:第 1 节简单介绍 NMT、PTM 和 PTNMT 的背景;第 2 节介绍 PTNMT 相关工作的分类对比;第 3 节总结 PTNMT 的优缺点并进行相关分析;第 4 节对 PTNMT 研究的未来进行展望;第 5 节总结本文内容。

1 背景

本节主要介绍 NMT、PTM 和 PTNMT 的概念,分析当前 NMT 和 PTM 的架构特点,最后阐述 PTNMT 的概念和当前面临的挑战。

1.1 NMT 的概念

NMT 指的是使用神经网络完成机器翻译的一种方法。近几年,随着深度学习技术的发展,基于端

到端表示学习的方法逐渐改变了自然语言处理的方式,NMT 在这种背景下应运而生。

NMT 模型包括编码端和解码端两个部分,其中编码端的作用是处理源语言并生成中间表示;解码端利用中间表示,使用自回归的方式(利用当前生成的词与上下文向量一同预测下一个词),逐词生成目标语言。NMT 模型的这种利用编码端和解码端的结构称为端到端结构,如图 1 所示。

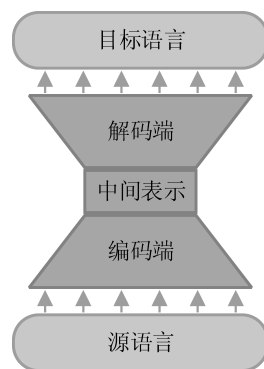


图 1 NMT 的端到端结构图

通过该结构,NMT 模型可以完成源语言到目标语言的转换。这种结构的一个优点是编码端和解码端是可替换的“插件”,从某种意义上讲,编码端和解码端是“低耦合”的。这些插件可以进行替换,为 PTNMT 提供了可能。在这种范式下,基于注意力机制实现的 Transformer 模型已经在 NMT 中占据了统治地位。

1.2 PTM 的概念

PTM 指的是使用大规模数据(单语数据、双语数据或单语和双语数据)在特定的预训练任务上训练得到的模型。近几年,随着自注意力神经网络 Transformer 的出现,许多高性能的 PTM 被提出,例如,仅利用 Transformer 编码端构建的 BERT^[5]、XLM^[6]等;仅利用 Transformer 解码端构建的 GPT^[7]等;以及利用 Transformer 编码端和解码端构建的 MASS^[8]、BART^[9]等。这些工作表明,PTM 可以提取出“通用表示知识”,这使得 PTM 能够在下游任务上取得先进的性能,同时可以避免从零开始训练新模型^[11]。PTM 强大的性能引起了自然语言处理领域各个方向研究人员的关注。

当前主流的 PTM 框架分为两个阶段,分别是预训练阶段和引入下游任务阶段,如图 2 所示。在预训练阶段,预训练策略包括使用不同的预训练数据、模型结构以及预训练任务;在引入下游任务阶

段,引入策略包括微调、Prompt 以及信息融合等方法。PTM 的预训练策略决定了 PTM 语言建模和语义信息抽取能力的强弱,而 PTM 引入下游任务的策略则决定了 PTM 在下游任务上保留通用表示知识的多少。因此,使用更优的预训练策略和引入下游任务策略可以更好地完成下游任务。

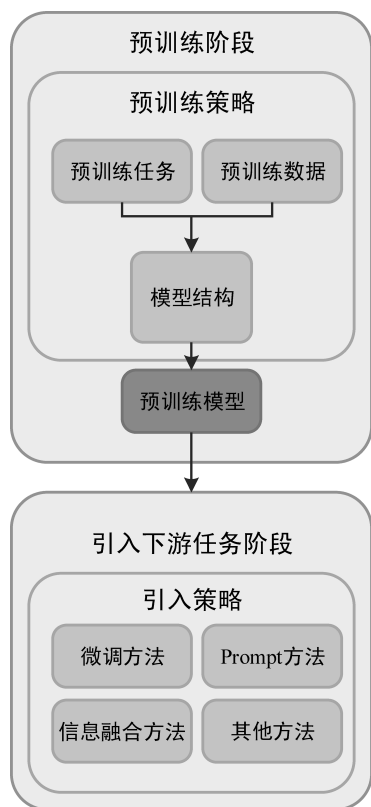


图2 PTM 相关方法的两个阶段

1.3 PTNMT 的概念和面临的挑战

1.3.1 PTNMT 的概念

NMT 模型端到端的结构虽然不需要人工参与进行特征的提取,但却带来了大量需要优化的参数。为了优化这些参数,往往需要大量的双语数据,并且机器翻译任务的复杂性较高,需要更多的双语数据,而现实中构建大规模标注双语数据是一个巨大的挑战,一方面是只有少部分语言对拥有高质量大规模数据,如中英、英法等;另一方面是当前的数据在领域上分布不均,现有的高资源数据一般集中在新闻、专利等领域。这会导致 NMT 模型训练不充分或难以训练,模型很难完成语言建模以及语义信息的抽取,极大地影响了 NMT 的性能。

相比之下,大部分语言的大规模无标签单语数据集较为容易构建。为了利用大规模无标签数据,

研究人员通常先在其上学习通用表示,而后在下游 NMT 任务上使用这些表示。最近的研究表明^[5-9],在大规模无标签数据上训练的 PTM 可以提取“通用表示知识”,并在许多自然语言处理任务上展现出强大性能,这引起了机器翻译研究人员的关注。研究人员引入 PTM 以实现性能更强大的 NMT 模型,这种方法被称为 PTNMT。

1.3.2 PTNMT 面临的挑战

虽然 PTNMT 与 NMT 相比具有三大优势,但实现 PTNMT 也面临着挑战,下面将展开分析。PTNMT 构建主要分为两个阶段,首先在大规模数据上完成 PTM 的训练,即 PTM 的预训练阶段;而后将 PTM 引入下游 NMT 任务,即 PTM 引入下游任务阶段。如前文所述,PTM 在预训练阶段使用的预训练策略包含三个方面:预训练数据、模型结构以及预训练任务,这三方面与下游 NMT 任务都存在不匹配的现象,并且这三方面不仅影响 PTM 自身的性能,同时在引入下游任务时也会对下游任务造成极大的影响^[10],因此如何实现高性能的 PTNMT 模型具体面临以下挑战,从以下三点进行分析。

(1) **PTM 与 NMT 模型的训练数据不同带来的影响** PTM 往往利用无标签的单语数据进行预训练,而下游 NMT 任务则使用双语数据进行训练,这会导致仅使用无标签数据训练的 PTM 无法直接构建 NMT 模型。同时,PTM 和 NMT 模型的训练数据也存在领域和问题不匹配的问题。除此之外,有研究工作^[14]表明:由于训练数据的不同,PTNMT 模型可能会产生复制翻译或幻觉翻译(生成非目标语言)的问题。一旦出现这样的问题,将会导致生成句子的语义流畅性大大下降,甚至背离机器翻译任务的初衷。

(2) **PTM 与 NMT 模型的模型结构不同带来的影响** 当前存在使用 Transformer 的编码端、解码端以及编码端和解码端构建的 PTM,而 NMT 是一个具有编码端和解码端的模型。这就可能会产生 PTM 与 NMT 模型结构的不匹配问题,进而导致 NMT 模型无法充分利用 PTM 的语言建模和语义信息抽取能力。

(3) **PTM 与 NMT 模型的训练任务不同带来的影响** PTM 的预训练任务往往是无监督任务,而下游 NMT 任务执行的是有监督的翻译任务,两者存在天然的差距。任务上的差距会导致:将 PTM 引入 NMT 时,随着模型的训练,PTM 出现灾难性遗

忘的情况(PTM 在执行下游任务时失去了通过预训练获得的能力)。

因此为了实现高性能的 PTNMT 模型,一方面需要设计 PTM 的预训练策略以适配下游 NMT 任务,从根本上消除预训练策略对下游 NMT 任务的影响;另一方面需要设计 PTM 的引入策略来缓解 PTM 的预训练策略与下游 NMT 任务不匹配带来的影响。

在本文的后续各节中将介绍当前存在的 PTNMT 工作,并分析这些工作的特点以及是如何应对上述挑战的。

2 PTNMT 方法分类

近几年涌现出许多 PTNMT 的相关工作,这些工作可以从 PTNMT 的构建、问题分析和应用入手分为以下四个类别:(1)PTNMT 中 PTM 的预训练策略;(2)PTNMT 中 PTM 的引入策略;(3)PTNMT 模型及数据分析;(4)PTNMT 在特定环境中的应用,如图 3 所示。其中,前两个类别分别从 PTNMT 构建的两个阶段对 PTNMT 方法进行详细分类介绍。

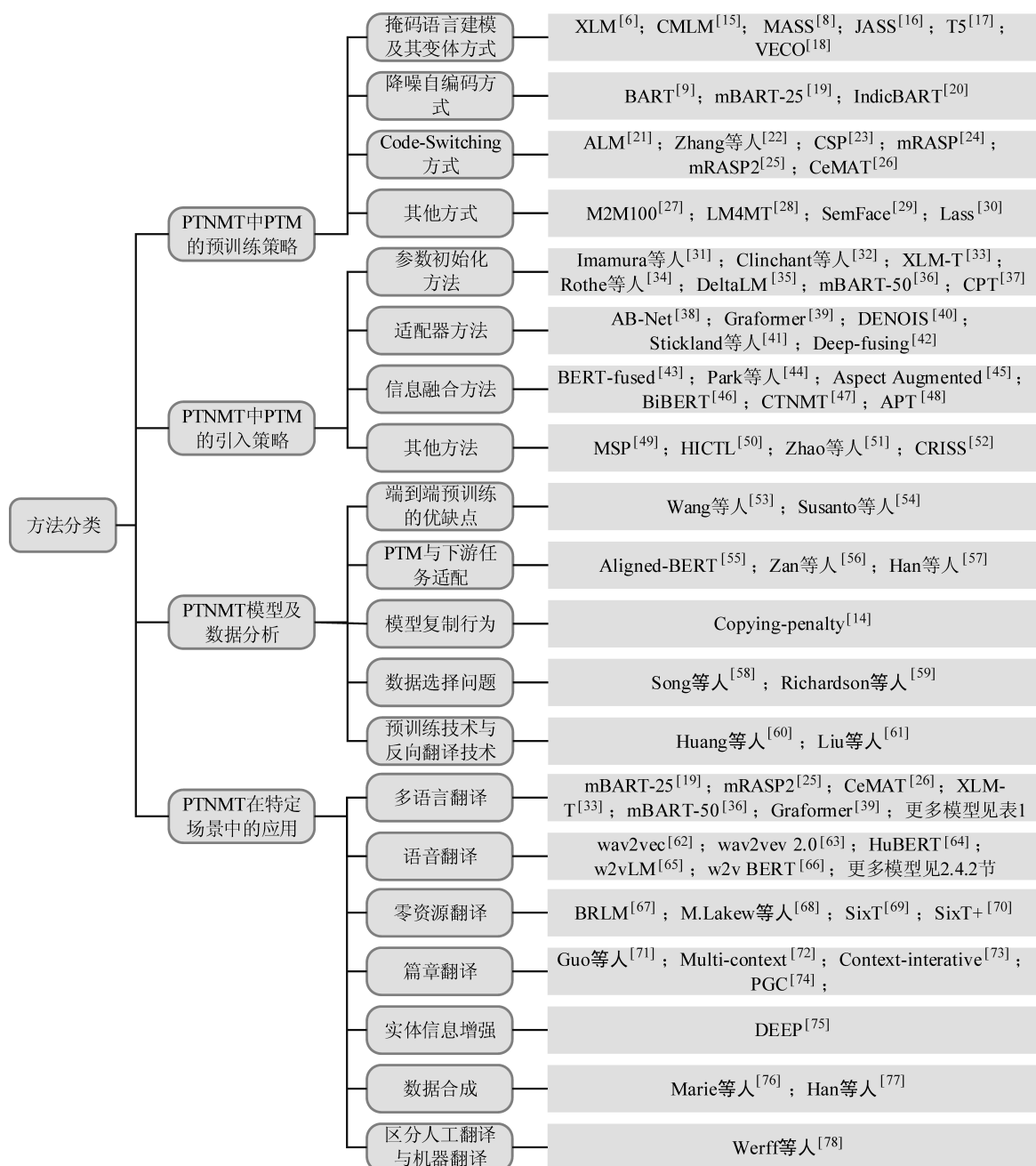


图 3 PTNMT 方法分类图

具体来说,PTNMT 中 PTM 的预训练策略指的是研究人员通过设计不同的预训练策略来实现专门用于 NMT 的 PTM;PTNMT 中 PTM 的引入策略指的是研究人员使用已有的性能强大的 PTM 通过特定的策略将其适配于下游 NMT 任务;PTNMT 模型及数据分析指的是研究人员对 PTNMT 模型的结构和使用不同数据的影响进行分析,以解释 PTNMT 模型的行为;PTNMT 在特定环境中的应用指的是研究人员将 PTNMT 模型应用于解决某些特定环境下的问题。

接下来本文将根据上述分类对 PTNMT 现有工作进行详细的总结和分析。

2.1 PTNMT 中 PTM 的预训练策略

PTNMT 构建的一个重要步骤是 PTM 的预

训练,与下游任务更为匹配的预训练策略能够更好地提升 PTNMT 模型的性能。本节介绍的工作是如何从零训练 PTM,而后将 PTM 应用于 NMT 中。

2.1.1 掩码语言建模及其变体方式

掩码语言建模(Masked Language Modeling, MLM)是 PTM 常见的一种预训练方式,Kenton 等人^[5]首次将该方式应用于 PTM 中以克服标准语言模型的单向建模缺点,并提出 BERT 模型。MLM 对输入句子进行掩码操作,将被掩码的词汇替换为统一的“[MASK]”标记,并训练模型通过未掩码词汇来预测掩码词汇,如图 4(a)所示。MLM 训练方式使 BERT 模型可以双向建模语言文本,更好地完成语言建模和语义信息抽取。

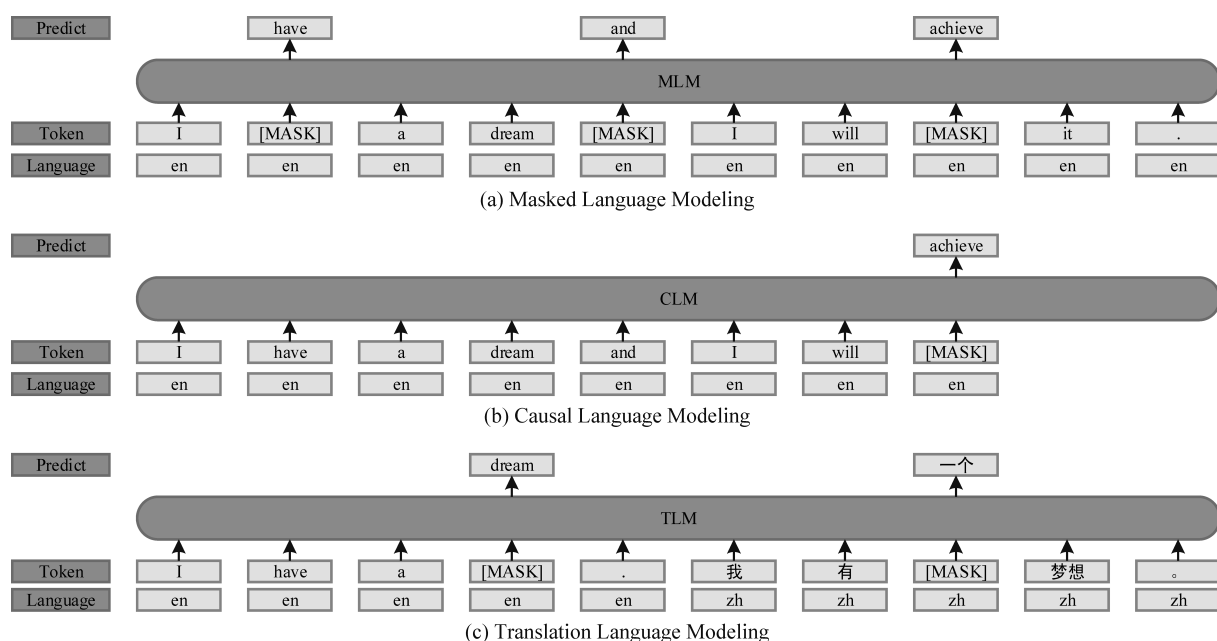


图 4 MLM、CLM 以及 TLM 图示

跨语言掩码语言建模 为了使 PTM 具有跨语言信息,Lample 等人^[6]提出 XLM 模型,该模型在多语言数据集上训练,除使用 MLM 预训练任务外,新增了因果语言建模(Causal Language Modeling, CLM)和翻译语言建模(Translation Language Modeling, TLM)两种预训练任务,CLM 和 TLM 方式如图 4(b)、图 4(c)所示。Lample 等人分别使用 CLM、MLM 或在 TLM 基础上使用 MLM 共三种任务训练模型。结果表明,MLM 预训练任务对无监督机器翻译和监督机器翻译都非常有效。同时 Lample 等人还发现:①添加更多的语言可以提高低资源语言的性能;②频繁地采样某种语言可以提

高这种语言的性能,但会损害其他语言的性能。

为了增强 PTM 对跨语言信息的学习能力,Ren 等人^[15]提出跨语言掩码语言模型 CMLM。Ren 等人首先构建跨语言翻译表,而后对源语言句子中出现在翻译表内的词汇进行掩码,并在输出时预测其在翻译表中的候选翻译。通过这种方式,Ren 等人将跨语言信息融入 PTM 中。Ren 等人使用 CMLM 模型作为翻译模型的编码器和解码器,显著地提高了无监督机器翻译的性能。

端到端掩码语言建模 虽然 XLM 模型能够提取跨语言信息,但其模型结构仅具有 Transformer 的编码端,而端到端 NMT 模型往往采用编码端-解

码端的结构,因而会产生模型结构不匹配的问题。为了解决该问题,Song 等人^[8]提出掩码端到端预训练模型 MASS,该模型在无监督机器翻译和低资源机器翻译上效果显著。Mao 等人^[16]在 MASS 的基础上引入语义信息,根据语言特性,将输入句子按照语义单元重新切分。在掩码过程中,不再随机掩码词汇,而是将划分好的语义单元作为最小单元进行掩码。除此之外,Mao 等人又增添恢复乱序语义单元的重排序任务。Mao 等人的实验表明,引入语义信息的 MLM 和重排序任务具有良好的互补性,能够大幅提升低资源机器翻译的性能。

Raffel 等人^[17]引入了一个统一的架构,将所有基于文本的语言问题转换为文本到文本的问题。在预训练时,Raffel 等人使用无监督 MLM 任务和监督的下游任务(包括机器翻译、文本摘要等任务),在端到端的架构上进行多任务学习,最后得到 T5 模型。为了在文本到文本格式中区分不同的下游任务,该模型为每一个下游任务添加了一个 Prompt。

模型模块化重组 Luo 等人^[18]认为,仅用 Transformer 的编码端训练用于理解任务的模型和用 Transformer 的编码端-解码端训练用于生成任务的模型,忽略了两个任务和模型框架之间的相关性。因此 Luo 等人提出 VECO 预训练方法,该方法

将 Transformer 拆分成多个子模块,并使用提出的两种预训练任务单序列掩码语言建模 (Inner-Sequence Masked Language Modeling, IS-MLM) 和交叉序列掩码语言建模 (Cross-Sequence Masked Language Modeling, CS-MLM) 进行训练,在下游任务使用时再根据需要将子模块组装起来。这种共享通用子模块的方式可以使不同模块相互促进,并且相比于仅用 Transformer 编码端的 PTM,该模型拥有交叉注意力参数,更有利于完成机器翻译任务。

2.1.2 降噪自编码方式

降噪自编码方式 Lewis 等人^[9]提出了一个基于端到端架构使用降噪自编码方式的预训练模型 BART,如图 5(c)所示。BART 的预训练过程为:①使用任意的噪声函数破坏输入文本;②训练模型来重建输入文本。Lewis 等人共设计了五种噪声破坏函数,分别是:①词汇掩码,替换为“[MASK]”;②词汇删除;③片段掩码,一段词汇替换为“[MASK]”;④句子排序,将乱序的句子排序;⑤文档旋转,识别文档的开头位置。由于 BART 模型在单语数据上进行训练,不具备其他语言的知识。因此,在执行 MT 任务时,Lewis 等人将 BART 模型视为 MT 系统的解码端,通过新增额外的编码端完成 MT 模型搭建。

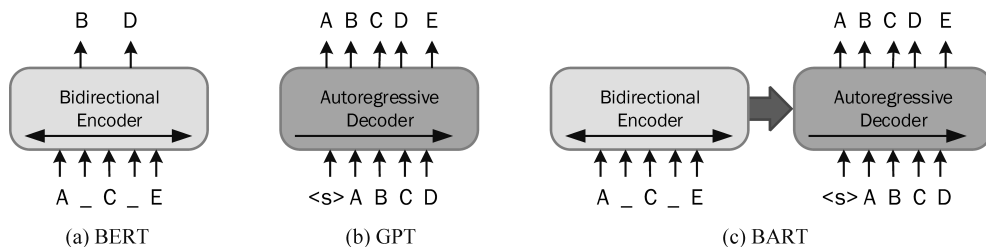


图5 BERT、GPT 以及 BART 的模型结构及输入输出图示^[9]

多语言降噪自编码方式 为了克服 BART 模型单语种的缺点,Liu 等人^[19]提出了 BART 的多语言版本模型 mBART。mBART 训练数据为 25 种语言的大规模单语语料库,训练过程与 BART 相似。由于 mBART 是端到端的多语言预训练模型,因此在执行下游 NMT 相关任务时直接用相关数据进行微调即可。Liu 等人在无监督 NMT、监督 NMT 和篇章 NMT 上评估了 mBART。实验结果表明,mBART 对低资源和中资源翻译有显著的效果提升,而当微调的双语数据是超过 2 500 万的高资源时性能略有下降,Liu 等人怀疑高资源数据的有监督训练会完全“清洗掉”预训练的权重。在多语

言语料库上训练的 mBART 具有强大的跨语言能力,对未出现在预训练语料库的语言进行微调也能够得到性能的提升。

针对特定语系语言 Dabre 等人^[20]利用印度文字之间的正交相似性设计了一个针对印度语系语言和英语的预训练模型 IndicBART。在训练过程中利用印度语系特性,Dabre 等人将这些印度语转换成单一文本。在 NMT 上的实验结果表明,IndicBART 可以有效提升印度语系的迁移学习能力,实现更高的 NMT 性能。Dabre 等人认为基于多语言微调和以某种语言为中心的单一文本模型能够实现良好的性能。

2.1.3 Code-Switching 方式

Code-Switching 方式 将输入句子中的部分词汇替换为另外一种语言的同义词汇并组成新的句子, 这种方式能够有效地使模型学习到跨语言信息, 如图 6 所示。

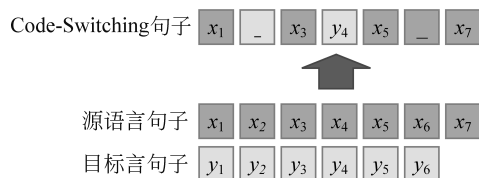


图 6 Code-Switching 图示

Yang 等人^[21]利用 Code-Switching 方式提出了一种新的跨语言预训练方法——交替语言建模 (Alternating Language Modeling, ALM)。与 TLM 方式相比, ALM 方式不再将不同语言的句子进行简单拼接, 而是使用 Code-Switching 方式, 用目标词汇替换源词汇, 构建新的句子。随后对构建的句子做随机掩码, 预训练任务的目标是预测新的句子中被掩码的词汇。Yang 等人希望通过这种方式捕获单词级和短语级丰富的跨语言信息。Zhang 等人^[22]将 Code-Switching 方式引入 XLM 和 MASS, 与 ALM 交替预测不同语言词汇的方式不同, Zhang 等人先对句子进行掩码, 而后对未掩码词汇做 Code-Switching。Zhang 等人分析, 通过这样的方式能够使模型学习到更多的对齐知识, 有利于低资源机器翻译。

端到端地使用 Code-Switching 方式 为了与下游 NMT 任务具有同样的模型架构, Yang 等人^[23]使用 Code-Switching 方式提出了一种基于端到端架构的预训练方法 CSP。与传统的 MLM 预训练方法不同, CSP 随机将输入句子的一些词汇替换为目标句子的翻译词汇。CSP 方法的预训练目标是预测被替换的词汇对应的源词汇。Yang 等人认为 CSP 方法能够显式地充分利用从源语言和目标语言单语语料库中提取的跨语言对齐信息来预训练 NMT 模型, 同时这种方法还能缓解由人工符号(例如“[MASK]”)引起的差异问题。

翻译任务作为预训练任务 为了更加适配下游任务, Lin 等人^[24]将翻译任务作为预训练任务, 在端到端架构上利用 Code-Switching 方式构建了预训练模型 mRASP。mRASP 的训练数据是并行的多语言数据, 在预训练阶段, 使用随机对齐替换技术, 使不同语言内具有相似含义的词汇和短语在表示空

间中更接近。相比于 mBART, mRASP 不仅可以提升低资源 MT 和中资源 MT 的性能, 还可以提升高资源 MT 的性能。mRASP 能够用于零样本翻译, 甚至可以提升从未在预训练语料库中出现的语言的翻译质量。Pan 等人^[25]在 mRASP 模型的基础上提出了 mRASP2 模型。相较于 mRASP 模型, mRASP2 模型的训练数据包括双语数据和单语数据, 并在编码端引入对比学习来缩小不同语言表示之间的差距。

双向解码器解码 与以往的单向解码器不同, Li 等人^[26]使用双向解码端构建了一个端到端预训练模型 CeMAT, 该模型可为自回归和非自回归 NMT 带来显著的性能提高。CeMAT 模型训练数据包括单语数据和双语数据, 训练目标是分别在编码端和解码端预测掩码词汇。Li 等人为了增强 CeMAT 模型的性能, 引入了两种简单的方法: ①与掩码结合的 Code-Switching; ②动态掩码。Li 等人的大量实验表明, CeMAT 在从低资源到极高资源的语言场景中都有性能的提高。

2.1.4 其他方式

除了上述的预训练方法分类外, 研究人员还尝试了其他的预训练方法。

混合专家系统 Fan 等人^[27]利用多专家混合系统思想, 设计了一个 100 种语言可以相互翻译的模型 M2M100。M2M100 模型设计了不同语言的路由结构, 并在网络的最后一层添加了语言特定的预测层。M2M100 通过密集缩放和特定语言的稀疏参数两种方法, 有效增加模型容量。

非端到端翻译模型 Wang 等人^[28]设计了一个非端到端的翻译模型 LM4MT, 在训练过程中将源语言句子和目标语言句子拼接在一起, 同时优化语言模型损失和翻译任务损失来完成训练, 最后实现的性能与基线性能相当。

语义接口 为了预训练交叉注意模块参数, Ren 等人^[29]在预训练的编码端和解码端之间定义语义接口, 提出了预训练方法 SemFace。Ren 等人在预训练阶段先完成编码端的训练, 而后再利用编码端和定义的接口完成解码端的训练。SemFace 模型在无监督和有监督 NMT 上的性能均表现良好。

子网络思想 Lin 等人^[30]认为由于参数干扰, 多语言 NMT 模型会在极高资源的语言对上出现性能下降的现象。为了缓解参数干扰问题, Lin 等人利用子网络思想提出了 LaSS 模型。LaSS 模型通

过学习每个语言对的语言特定子网络来对抗参数干扰。实验表明,LaSS 方法能够快速适应新的语言对,在零资源翻译方面表现出很强的泛化能力。

2.2 PTNMT 中 PTM 的引入策略

从零训练 PTM 代价很大,因此有效利用已有

PTM 并将其引入 NMT 具有很大的价值。本节介绍的工作是将已有的 PTM 引入 NMT 时的策略,三种引入策略的简单描述如图 7 所示。优秀的引入策略一方面可以避免 PTNMT 的灾难性遗忘问题,另一方面可以更好地让 PTM 适配下游任务。在不加特殊说明的情况下,NMT 模型结构都是以 Transformer 为基础。

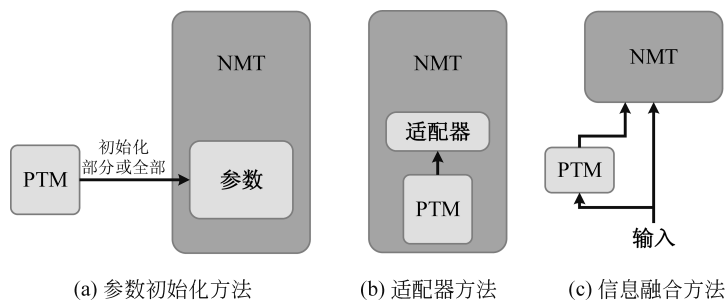


图 7 三种将 PTM 引入 NMT 的策略图示

2.2.1 参数初始化方法

PTM 仅初始化编码端 基于 Transformer 模型的 PTM 在 NMT 任务上拥有天然的结构优势。Imamura 等人^[31]使用 BERT 模型初始化 NMT 模型的编码端。为了对整个模型进行合适的训练,Imamura 等人采用了两阶段优化,首先通过冻结 BERT 模型只训练解码端参数,然后对整个模型进行微调。该方法有利于低资源翻译。

Clinchant 等人^[32]同样使用 BERT 模型初始化 NMT 的编码端,但侧重点是探索 BERT 模型不同位置参数的作用。Clinchant 等人分析了引入词嵌入层参数、引入所有参数并微调以及引入所有参数并冻结这三种方式对 NMT 模型的翻译性能的影响。实验结果表明:①冻结 BERT 的所有参数,将 BERT 的输出直接作为编码结果,这种方式得到的编码结果不足以执行翻译任务;②引入 BERT 所有参数并微调的方式性能最优。除此之外,还探索了使用不同的单语数据训练 BERT 模型对下游翻译任务造成的影响。在源语言数据集上训练 BERT 是有益的,这也说明 BERT 模型能够为 NMT 模型的编码端提供更好的初始化。

PTM 初始化编码端和解码端 Ma 等人^[33]利用多语言预训练模型 XLM-R 搭建多语言机器翻译模型 XLM-T。与前面研究人员的初始化方法不同,Ma 等人在初始化 NMT 模型时不仅初始化编码端,同时还初始化解码端。由于 XLM-R 模型是以 Transformer 模型的编码端为基础搭建的,因此在初始化 NMT 模型解码端时无法直接完成对交叉注

意力参数的初始化。Ma 等人通过几种方式来解决这个问题,包括:①共享交叉注意力参数与自注意力参数;②随机初始化交叉注意力参数。随后 Ma 等人使用多语言并行数据,对 XLM-T 模型进行微调。XLM-T 模型在多语言机器翻译上有着显著的改进,即使与反向翻译的方法相比,该模型也是有效的。

Rothe 等人^[34]利用 BERT、GPT-2、RoBERTa 三种模型对端到端模型进行初始化。初始化方法包括仅初始化编码端、仅初始化解码端以及初始化编码端和解码端这三种方式。在下游 NMT 任务上, Rothe 等人利用上述初始化方式进行实验,展现出不同策略的优缺点。

PTM 初始化交叉注意力参数 为了解决基于 Transformer 编码端构建的 PTM 无法直接初始化交叉注意力参数问题, Ma 等人^[35]提出 DeltaLM。DeltaLM 使用 InfoXLM 作为进行初始化的 PTM,对于编码端可以直接使用 InfoXLM 进行初始化,而解码端则使用交替初始化的方式,即使用 InfoXLM 的奇数层初始化自注意力和底部前馈网络的参数,使用相应的偶数层初始化交叉注意力和顶部前馈网络参数。完成初始化后继续对 DeltaLM 进行预训练,预训练任务包括:①单语片段掩码任务;②双语片段掩码任务。其中双语片段掩码任务将源语言句子和目标语言句子拼接在一起进行训练,目的是提高模型的跨语言能力。DeltaLM 在多语言机器翻译上有着显著的性能提高。

继续预训练方法 与前人使用的 PTM 类型不

同, Tang 等人^[36]使用端到端的 PTM 模型 mBART。Tang 等人首先创建了包含 50 种语言, 且涵盖低、中、高资源语言的数据集。随后 Tang 等人使用 mBART 初始化模型并在创建的数据集上继续预训练, 成功地将 mBART 支持的语言从 25 种扩展到 50 种。最后使用多语言并行数据对扩展后的模型进行翻译任务的训练。Tang 等人的方法实现了一个支持 50 种语言的机器翻译模型, 在低资源机器翻译上有着显著的性能提升, 且在扩展过程中对原有语言翻译没有损失性能。

Liu 等人^[37]也使用 mBART 初始化模型, 处理的任务是低资源机器翻译。虽然 mBART 对低资源机器翻译性能有着显著的改进, 但是, 当翻译在预训练数据中未出现的语言时, 它的性能将受到很大限制。Liu 等人提出了一个继续预训练架构 CPT, 以有效地使 mBART 适应未知的语言。具体做法是使用未知语言的单语语料库构造出未知语言和已知语言的混合语言文本, 然后对 mBART 继续预训练, 以重构未知语言文本。结果表明, Liu 等人的方法与 mBART 原始方法相比, 对未知语言的低资源机器翻译有着显著的性能提高。

2.2.2 适配器方法

由于 PTM 的预训练任务始终与下游 NMT 任务存在天然的差距, 为了减少因上下游任务不同带来的灾难性遗忘问题, 研究人员在设计模型时引入神经网络组件(适配器)来缓解该问题带来的影响, 此类方法称为适配器方法。

层级别适配器方法 Guo 等人^[38]提出 AB-Net 方法, 使用 BERT 模型初始化 NMT 模型的编码端和解码端, 并在 BERT 每一层后加入前馈网络作为适配器模块。随后使用双语数据进行微调, 但在微调阶段只对适配器模块的参数进行微调, 其他部分参数通过冻结保持不变。Guo 等人认为该方法可以联合利用编码端和解码端 BERT 模型中的信息, 还能避免产生灾难性遗忘问题。该方法的另一特点是其可以应用于自回归和非自回归的 NMT 任务上。由于通常 PTM 层数较多, 将适配器插入所有层将导致大量新参数, 故 Guo 等人^[79]后续又提出 Adaptive Adapters, 在该方法中引入了隐变量来决定是否在每一层中使用适配器。通过这种方式, PTNMT 模型能够自动确定要使用哪些适配器, 极大地提高了参数效率和解码速度。

网络级别适配器方法 Sun 等人^[39]分别使用 mBERT 和 mGPT 作为 NMT 模型编码端和解码端

的信息提取网络, 随后分别在 mBERT 和 mGPT 的后面增加 Transformer 的编码端和解码端结构, 进而形成 Graformer 模型。Graformer 模型在微调阶段冻结 mGPT 模型的参数, 微调 mBERT 模型和适配器模块参数。该方法一方面化解了预训练模型没有交叉注意力模块的问题, 另一方面可实现多语言机器翻译。

语言专用适配器 Üstün 等人^[40]使用 mBART 模型提出 DENOIS 方法, 在编码端和解码端的顶部对每一种语言添加一层适配器层, 并采用两阶段的训练方式。第一阶段在单语数据上只训练适配器层参数, 第二阶段在双语数据上只训练交叉注意力参数。对于没有双语数据的语言对, 可以仅在第一阶段训练后利用其他语言的适配器层实现无监督机器翻译。

分析适配器位置影响 Stickland 等人^[41]研究了 BART 类模型在 MT 任务上进行微调时冻结参数策略以及添加新参数位置的影响。得到的结论是: ①对于 BART 模型, 要冻结编码端和解码端参数, 并在 BART 编码端之前引入新的编码端, 在 BART 编码端之后添加适配器层; ②对于 mBART 模型, 只需要冻结 mBART 的解码端中除交叉注意力以外的参数, 并在 mBART 解码端后添加适配器层。两个结论的共同点是, 对生成端的参数冻结能够取得更好的效果。

多层融合适配器 为了缓解 PTM 优化目标与 NMT 优化目标差异性带来的影响, Weng 等人^[43]提出一种将 PTM 表示深度融合到 NMT 的新架构的方法。具体而言, Weng 等人使用 mBERT 模型初始化 Transformer 的编码端, 并采用分层协调结构来协调 PTM 和 NMT 解码器的学习。同时, Weng 等人引入了一种分区多任务方法来微调 PTM 的参数, 通过逐步学习的特定表示来缩小 PTM 和 NMT 之前的差距。

2.2.3 信息融合方法

信息融合方法 是将 PTM 作为一个独立的网络参与 NMT 模型的计算。与前两种使用策略不同, 信息融合方法不会直接向 NMT 模型提供参数, 而是向 NMT 模型提供 PTM 经过处理得到的信息, 这是一种能够实现 PTM 与 NMT 解耦的方法。

层交互的信息融合方法 Zhu 等人^[43]为了更好地在 NMT 中使用 BERT 模型, 提出了 BERT-fused 模型。该方法首先使用 BERT 提取源语言句子的表示, 然后通过注意力机制将提取的表示与 NMT 模型的编码端和解码端的每一层进行融合。

BERT-fused 模型在监督、半监督和无监督机器翻译任务上进行实验,与 Transformer 基线相比性能都有提高。

Park 等人^[44]探讨了如何有效利用 PTM 的中间输出来完成语言生成任务。该方法将输入句子经 NMT 编码端计算的结果与经 PTM 计算的结果进行融合,得到最终的上下文表示。该方法的特殊之处在于,PTM 的注意力机制中使用的查询信息来自于 NMT 编码端,相当于对 PTM 使用了同种语言的交叉注意力机制。Park 等人也对不同 PTM 进行融合的效果进行了实验。

语义嵌入的信息融合方法 Shavarani 等人^[45]认为将 PTM 引入 NMT 时,训练过程是脆弱且不可靠的,NMT 模型很难通过微调 PTM 的方法来获得增益。因此,Shavarani 等人提出了一种基于语义信息提取的方法 Aspect Augmented 来利用 BERT 模型,通过新增语义嵌入层从 BERT 提取语义信息,再经过整合网络将语义信息与 NMT 词嵌入信息进行融合完成模型的搭建。这种方法有助于 NMT 在各种训练环境中更好地泛化,并且与传统的基于 Transformer 的 NMT 相比,没有增加训练难度。

级联式信息融合方法 为了研究不同 PTM 对 PTNMT 翻译性能的影响,Xu 等人^[46]使用在不同结构和不同数据上训练的 PTM 进行实验。最后 Xu 等人发现将特定双语数据训练的 BiBERT 的输出作为 NMT 编码器的输入,可以获得显著的翻译性能提升。此外 Xu 等人还提出了一种随机层选择方法和双向翻译模型的概念,以确保充分利用上下文信息。

知识提取的信息融合方法 为了缓解 PTM 应用于高资源机器翻译时出现的灾难性遗忘问题带来的影响,Yang 等人^[47]提出了 CTNMT 框架。CTNMT 框架包括三种技术:①渐进蒸馏,确保 NMT 模型能够保留 PTM 的知识;②动态选择门,缓解 PTM 出现灾难性遗忘现象;③动态调整学习率策略。Yang 等人实验检验了 BERT 和 GPT 模型应用于 CTNMT 框架的效果,发现双向信息建模的 BERT 效果更好。

Weng 等人^[48]提出了一种将 PTM 信息用于 NMT 编码端和解码端的架构——APT 架构。APT 架构主要包括动态融合机制和知识提取范式。动态融合机制将 PTM 的知识融合到 NMT 的编码端,知识提取范式将 PTM 的知识融合到 NMT 的

解码端。Weng 等人研究了在 NMT 的编码端和解码端使用 BERT 和 GPT 模型不同组合信息融合的效果,最终发现在编码端使用 BERT,解码端使用 GPT 得到的 NMT 模型性能最好。

2.2.4 其他方法

基于 Prompt 微调方法 还有研究人员使用基于 Prompt 的微调方法,Tan 等人^[49]提出多阶段微调方法 MSP。为了使 PTM 更好地进行翻译,Tan 等人将翻译过程分为三个独立的阶段:编码阶段、重新编码阶段和解码阶段。在每个阶段,对 PTM 独立应用不同的连续提示,以使 PTM 更好地适应翻译任务。该方法选择的 PTM 为 GPT 模型,实现的翻译模型在低资源机器翻译上有性能的提高,但是在高资源机器翻译上性能不如基线 Transformer 模型。

对比学习 为了加强 PTM 的跨语言信息的学习能力,Wei 等人^[50]提出了层次对比学习方法 HICTL。该方法在 XLM-R 模型上继续预训练,通过对比学习实现了:①学习一种或多种语言对齐句子的通用表示;②通过构建正负例单词方式,将不同语言相同含义的单词距离拉近,不同含义的单词距离拉远。

课程学习 NMT 的性能很大程度依赖训练数据,然而高质量的训练数据难以获得。Zhao 等人^[51]提出基于 PTM 的课程学习训练 NMT 模型的方法。在训练时,首先使用原始数据训练得到一个预训练翻译模型,而后从原始数据中选择有影响的数据样本,学习改进预训练翻译模型的课程。在多个数据集上的实验表明,在原始数据集上训练收敛时,使用该方法可以进一步提高 NMT 模型的性能。

自监督跨语言训练架构 Tran 等人^[52]发现通过使用预训练模型输出挖掘的句子对端到端模型进行训练,可以进一步提升跨语言对齐效果。Tran 等人提出一种新的方法——基于迭代的自监督跨语言训练,即 CRISS,其中挖掘和训练过程被反复迭代。CRISS 同时也提高了跨语言对齐和翻译能力。

2.3 PTNMT 模型及数据分析

端到端预训练的优缺点 Wang 等人^[53]对端到端 PTM 的代表 mBART 引入 NMT 进行了研究,经过分析后发现 mBART 模型对 NMT 作用具有两面性:一方面,它有助于 NMT 模型产生更多样化的翻译,并减少与充分性相关的翻译错误。另一方面,端到端预训练和 NMT 微调之间的差异限制了

翻译质量。为了解决这些问题, Wang 等人提出了两个简单的策略, 分别是领域预训练和输入适应, 以弥补领域和客观的差异。Susanto 等人^[54]也对 mBART 进行分析, 分析的重点是将 mBART 进行多语言微调后对各种翻译任务的有效性, 包括低资源、多模态及跨领域翻译。Susanto 等人发现 mBART 能够成为构建多语言翻译系统的良好起点。最后还将基于 mBART 的单模态方法与多模态强基线进行了比较。

PTM 与下游任务适配 尽管 PTM 在其他任务中取得了巨大的成功, 但在 NMT 中, 如 BERT 之类与 NMT 编码端架构极其相似的 PTM 却不能直接取代编码端, 这引起了研究人员的兴趣。Vázquez 等人^[55]使用平均余弦相似性、上下文度量和表示相似性度量来比较两者创建的嵌入空间, 并发现 BERT 与 NMT 编码端提取的表示明显不同。为解决该问题, Vázquez 等人提出 Aligned-BERT 方法, 使用显式对齐和微调的方法将两者的嵌入空间拉近。

PTM 已经成功应用于低资源 NMT。然而 PTNMT 很难在富资源 NMT 上取得与随机初始化(Random Initialization, RI)对等的性能, 甚至有些时候 PTNMT 的性能更差。Zan 等人^[56]分析 PTM 与 RI 的互补性, 并发现: ①PTM 可以实现比 RI 更平坦的损失, 虽然不能提高准确性, 但是可以提高模型的泛化能力; ②与 RI 相比, PTM 并不会提高词汇选择的可信度, 而是通过更平滑的词汇概率曲线, 增加词汇选择的多样性。基于以上的发现, Zan 等人利用两者的互补性设计了模型融合算法, 该算法可以拉近 PTM 和 RI 的距离。

Han 等人^[57]针对超大 PTM 与较小 PTM 在下游 NMT 任务中的表现进行了研究, 发现: ①在不同领域的 NMT 上, 超大 PTM 与较小 PTM 产生的效果是不同的, 在较小规模的商用汽车数据集上超大 PTM 更优, 而在规模较大的临床数据集上超大 PTM 便失去了优势; ②使用同一模型处理相同任务时, 不同评价指标的结果也不同, 在有些评价指标上该模型性能更好, 而在有些评价指标上该模型性能不佳。

模型复制行为 Liu 等人^[14]发现了 PTNMT 的一个重要的副作用, 即模型复制行为, 这是因为 PTM 的预训练任务和 NMT 的训练任务存在差异。PTM 的目标是学习预测掩码词汇并复制其他未掩码词汇, 因此用 PTM 初始化 NMT 将会产生复制

行为。Liu 等人通过设置重复率进行判定, 得到 PTNMT 模型比标准 NMT 模型具有更大的复制率的结论。为此 Liu 等人提出一种简单的惩罚方式 Copying-penalty 来控制解码过程中的复制行为。

数据选择问题 使用大规模单语语料库训练的端到端预训练模型可以提高低资源机器翻译的性能。但是, 某些语言并不具备大规模单语语料库。为此, Song 等人^[58]利用相似语言作为辅助语言与原始语言数据混合来弥补原始语言单语语料库的不足。为了进一步提高两种语言的相似性, Song 等人还使用了语言映射对齐的办法。除此之外, Song 等人还进行了数据选择, 将下游任务领域的数据作为训练数据。Richardson 等人^[59]不仅分析语言相似性带来的影响, 还分析其与预训练数据集规模、下游数据集规模三者对低资源机器翻译中性能的影响, 以及三者相互作用的方式。Richardson 等人发现更多的数据对双语系统有益, 同时三个因素之间存在统计性显著的交互作用, 这降低了大规模预训练数据集对某些下游任务的有效性。

预训练技术与反向翻译技术的对比 使用 PTM 实现 NMT 和使用反向翻译扩充数据是两种使用最广泛的单技术。Huang 等人^[60]使用 BERT-fused 模型与反向翻译进行对比, 发现: ①与反向翻译相比, 使用 BERT 实现的 NMT 模型在词法和句法的准确性方面有优势; ②BERT 可以提高 Transformer 在涉及语义和语用任务中的性能; ③在大规模数据集上训练的 PTM 可能会引入社会偏见, 影响翻译的公平性。Liu 等人^[61]从模型结构角度分析两者的作用, 发现预训练技术与反向翻译技术具有互补性, 预训练技术能够增强 NMT 模型的编码端, 而反向翻译可以增强 NMT 模型的解码端。

2.4 PTNMT 在特定场景中的应用

研究人员除了将 PTNMT 用在传统的翻译任务上, 还用其来解决一些特定场景下的问题。

2.4.1 多语言翻译

由于自然语言种类众多, 若为每一对语言都单独构建机器翻译模型, 则代价较大且效率较低。为此研究人员开始研究多语言机器翻译模型, 但同时多语言机器翻译模型也面临着因语种较多而难以保证每一个翻译方向性能的问题。

当前存在一些多语言 PTM^[19, 25, 27], 这类模型学习了不同语言的通用表示知识, 这种特性可以帮助其构建多语言机器翻译模型, 且效果很好。以 Liu

等人^[19]、Dabre 等人^[20]、Pan 等人^[25]、Li 等人^[26]以及 Fan 等人^[27]为代表的研究人员,通过设计 PTM 的预训练策略构建多语言 PTNMT;以 Ma 等人^[33,35]、Tang 等人^[36]、Sun 等人^[39]以及 Tan 等人^[49]为代表的研究人员,在已有的 PTM 基础上,通过设计 PTM 的引入策略实现多语言 PTNMT;这些工作的相关介绍在本文的 2.1、2.2、2.3 节中已给出,同时表 1 中“多语言”表项指出多语言 PTNMT 工作,这里不再赘述。

2.4.2 语音翻译

语音翻译根据生成目标的模态可以分为语音到语音翻译和语音到文本翻译,本节主要关注语音到文本的翻译。

语音翻译中使用 PTM 最常见的方法是,利用语音识别 PTM 的编码器和机器翻译 PTM 的解码器来初始化语音翻译模型相应部分的参数^[80-82]。由于在语音翻译模型中,编码器既要完成语音转录、生成语义表示,又要学习跨语言信息,因此编码器的训练难度较高。Bansal^[80]等人在高资源的语音识别数据上进行预训练,而后微调语音翻译模型参数来提升性能。Wang 等人^[83]提出一种课程预训练方法,在预训练阶段让编码器学习到语法和语义信息。还有一些研究人员从提升解码器性能的角度切入。Alinejad 等人^[84]提出一种对抗性正则化方法使语音识别和机器翻译模型的编码器表示更接近。为了降低解码器的训练难度,Dong 等人^[85]提出了连续解码策略,Liu 等人^[86]提出了一种知识提取方法。还有一些工作跳出编码端和解码端对语音翻译模型性能影响的限制,如 B́ansal 等人^[87]通过改进预训练方法和多任务学习手段改进了语音翻译模型;Stoian 等人^[87]分析得出以下结论:使用大量无关语言数据进行预训练,比仅使用少量相关语言数据进行预训练的方式,模型性能要好得多。为了降低预训练与微调之间的差距,Wang 等人^[88]提出了一种联合编码网络的训练方式,在预训练阶段和微调阶段保持模型中的每一部分能力一致,并使用预训练注意力模块来弥补差距。

上述工作仅使用了语音识别或机器翻译的数据,语音到文本翻译的并行数据依然有限,因此许多工作^[89-92]试图使用大规模未标记的语音或文本数据训练语音翻译模型。为了降低自监督地进行语音表示学习的难度,一些工作利用未标记的语音数据对音频信号进行建模,如 wav2vec^[62]、wav2vec 2.0^[63]、HuBERT^[64]、wavLM^[65]和 wav BERT^[66]等。

为了更好地对语音信号建模,一类工作^[92-94]试图利用语音 PTM 作为特征提取器来表示语音信号,在预训练完成后,这类方法直接将语音 PTM 产生的表示输入到语音翻译模型的编码端中。另一类工作^[89,91]则是在语音翻译模型的编码端直接以自监督的方式进行学习。Chen 等人^[91]提出了一种简单技术,仅在语音侧以自监督的方式训练语音编码器。Wang 等人^[89]通过在 wav2vec 2.0 编码器上添加随机初始化的解码器模型,构建了一个序列到序列的模型。此外,Dong 等人^[95]利用 BERT 来强制上层编码器在没有额外数据的情况下尽可能多地产生语义信息来构建新模型。Le 等人^[94]对多语言语音翻译进行研究,发现特定语言的适配器可以增强模型对每一对语言的理解能力。

上述的许多预训练方法因只对一种模态进行建模而受到限制。为了打破这一限制,语音翻译任务需要对语音和文本进行联合建模,学习统一的表示。Zheng 等人^[96]通过在全类语料库中联合学习语音和文本的统一表示来构建语音和文本的掩码语言模型。Ye 等人^[97]提出一种跨语音和文本模态的模型,该模型将语音和文本作为输入,并输出语音识别和文本翻译结果。Xu 等人^[98]提出了一种用于语音翻译的联合语音和文本编码的方法。Bapna 等人^[99]构建了一个具有在无标记文本和无标记语音上学习的编码器。Ao 等人^[100]提出了一个语音和文本统一模态的模型 SpeechT5,该模型实现了自监督语音和文本统一表示的端到端的预训练方式,支持语音识别和语音翻译等生成任务。Li 等人^[101]提出一种只微调 PTM 的标准化层和注意力层参数的方式,实现了将语音 PTM 编码器和文本 PTM 解码器结合以构建多语言语音翻译模型。

2.4.3 零资源翻译场景

在零资源翻译任务中,由于没有源语言到目标语言的双语数据,因此需要引入枢纽语言。例如,为实现闽南语与英语的互译,可以选择汉语作为枢纽语言。枢纽语言需要与源语言和目标语言都存在双语数据,且尽可能地与源语言相似。

Ji 等人^[67]提出新的方法 BRLM,在源语言和枢纽语言的单语和双语数据上训练了 PTM,并使用词对齐的方法完成源语言和枢纽语言的语义共享。随后利用 PTM 搭建 NMT 模型并使用枢纽语言到目标语言的双语数据进行训练。最后直接使用源语言在 NMT 模型上进行翻译,可以实现零资源翻译。

Lakew 等人^[68]的方法不要求枢纽语言与源语

言和目标语言存在双语数据。训练过程分为三个阶段:①使用包含目标语言的 PTM 初始化 NMT 模型;②使用目标语言单语数据增强 NMT 编码端;③将 NMT 模型的翻译方向转化为源语言到目标语言。其中后两个步骤构成一个学习周期。

Chen 等人^[69]提出了一个简单跨语言迁移模型 SixT,该模型使用仅包含一对并行语言对的数据集在 XLM-R 模型上进行训练。具体是使用 XLM-R 模型初始化整个编码端参数和解码端的词嵌入层参数,然后使用两阶段训练策略训练编码端和解码端。该模型在零资源翻译上性能显著优于 mBART。后来 Chen 等人^[70]在 SixT 的基础上提出“SixT+”模型,这是一个多语言到英语的 NMT 模型,支持 100 种源语言,但只需要使用 6 种源语言到英语的双语数据进行训练。其训练过程与 SixT 相同。

2.4.4 篇章翻译

由于 BERT 的预训练包括下一句子预测,Guo 等人^[71]利用此属性来改进篇章翻译。在 Guo 等人提出的模型中,BERT 作为上下文编码器实现文档级上下文信息提取,然后将该信息输入到翻译模型的编码端和解码端中,模型架构与 BERT-fused 相似。

Donato 等人^[72]提出一种使用 PTM 提取并融合上下文信息的架构 Multi-context,通过合并多个预处理篇章的上下文信息,使 Transformer 模型适应句子级序列。该架构使用了两种 PTM 提取上下文信息,分别是 BERT 和 PEGASUS,Donato 等人认为两个 PTM 可以为篇章翻译模型提供不同的上下文信号。Donato 等人还发现:①使用 PTM 提取上下文信息显著提高了模型性能;②充足的并行数据资源对学习使用文档上下文信息至关重要;③使用多个上下文信息的联合表示优于只使用一种上下文信息;④源语境比目标语境更需要上下文信息。

与前面两个工作使用已有 PTM 的方式不同,Yang 等人^[73]设计了一个专门用于篇章翻译的预训练策略 Context-interactive,通过设计三种特殊的预训练任务完成对上下文信息的建模。该模型通过跨句子生成中间句子,捕获文档中的跨句子依赖信息,并通过跨句子翻译更好地提取上下文信息。

Chen 等人^[74]提出一种全局上下文训练方法 PGC 来处理篇章翻译任务。PGC 的预训练任务包括两个训练目标:①重构带噪声的句子;②使用某

一句子的左右相邻句子生成该句子。模型包括全局上下文编码器、句子编码器和解码器,其结构与传统的 NMT 模型相似。

2.4.5 实体信息增强

机器翻译模型通常会对训练语料库中罕见的命名实体产生较差的翻译结果。Hu 等人^[75]提出 DEEP 方法,这是一种降噪的实体预训练方法,它利用大量单语数据和知识库来提高句子中命名实体翻译的准确性。此外,Hu 等人还使用多任务学习策略,即在实体增强的单语数据和并行数据上微调 PTNMT 模型。

2.4.6 数据合成

在 NMT 中,合成数据通常会提高模型的性能,特别是在低资源语言对和稀缺数据领域。

Marie 等人^[76]提出一种新的方法,利用特定领域中的少量单语数据生成大规模合成并行数据。具体做法是:首先使用特定领域中的少量单语数据对 GPT-2 进行微调,并得到大量的单语合成数据;接下来进行正向翻译和反向翻译,以合成域内并行数据。

Han 等人^[77]利用 PTM 的数据合成办法搭建了无监督机器翻译架构。Han 等人的方法包括三步:①使用 PTM 模型进行零资源翻译,为少部分句子生成翻译;②使用零资源翻译的生成数据作为小样本学习数据,采样合成更大的数据集;③舍弃部分数据,然后进行微调提取。其中用来做零资源翻译的模型是 GPT-3。

2.4.7 区分人工翻译与机器翻译

Werff 等人^[78]利用 PTM 来自动区分人工翻译和机器翻译的文本,并且在该项工作中有以下发现:①选择性能更强的 PTM 取得的性能会更好,不同 PTM 之间的性能差距很大;②只有在数据不足的情况下,才适合添加翻译的合成数据;③对文档进行分类和规范化,可以得到显著的性能改进。

2.5 PTNMT 相关工作总结

通过对当前 PTNMT 工作进行分析,本文发现,在大多数情况下,具有以下特点的 PTM 对下游 NMT 任务有较大帮助:

(1) 包含 Transformer 解码端结构的 PTM。这类 PTM 拥有文本生成能力,与翻译任务契合;

(2) 在预训练阶段使用了大量的数据(单语数据或双语数据)的 PTM。这类 PTM 会从大规模语料中学习语法、语义知识,且随着数据量的增大,

学习到的知识也变得丰富；

(3) 拥有大量参数的 PTM。这类 PTM 通过众多的参数拟合语法、语义空间,当其被迁移至机器翻译任务时,可以使 PTNMT 更好地理解语言。

但当前 PTNMT 工作不存在统一的范式,仍有很多问题亟待解决。PTNMT 工作详细的分析在第 3 节中介绍,未来研究方向在第 4 节中介绍。

前文所述工作的总结可参见表 1。其中,表项“PTM 为端到端”指的是构建 PTNMT 模型时使用的 PTM 是否为端到端的 PTM 模型;表项“其他 NTMT 实验”中的四个小项依次指的是“多语言机器翻译”“低资源机器翻译”“无监督机器翻译”以及该工作中的其他相关实验;最后四列是 NMT 中常用的基准数据集测试结果,因语音翻译实验所使用的基准数据集与 NMT 不同,故表 1 没有将语音翻译模型列入其中。

表 1 PTNMT 相关工作的实验类别及双语机器翻译性能对比

类别	模型	PTM 为 端到端	其他 NMT 实验				WMT14/%		WMT16/%	
			多语言	低资源	无监督	其他实验	EN-DE	DE-EN	EN-RO	RO-EN
PTNMT 中 PTM 的预训练策略	XLM ^[6]	否	—	—	√	—	26.40	34.30	33.30	31.80
	CMLM ^[15]	否	—	—	√	—	—	—	—	—
	MASS ^[8]	是	—	√	√	—	28.30	35.20	35.20	33.10
	JASS ^[16]	是	—	√	—	—	—	—	—	—
	T5 ^[17]	是	—	—	—	—	30.90	—	28.00	—
	VECO ^[18]	否	—	—	—	—	31.50	—	—	—
	BART ^[9]	是	—	—	—	—	—	—	—	37.96
	mBART-25 ^[19]	是	√	√	√	—	29.80	34.00	35.00	30.50
	IndicBART ^[20]	是	√	√	—	—	—	—	—	—
	ALM ^[21]	否	—	—	—	—	29.22	—	—	—
	Zhang 等人 ^[22]	都包含	—	√	—	—	—	—	—	—
	CSP ^[23]	是	—	—	√	—	28.90	—	—	—
	mRASP ^[24]	是	—	√	—	—	30.30	34.40	39.00	37.40
	mRASP2 ^[25]	是	√	√	√	零样本	—	—	38.00	39.10
	CeMAT ^[26]	是	√	√	—	—	30.00	—	38.00	37.10
	M2M100 ^[27]	是	√	√	—	零样本	—	—	—	—
	LM4MT ^[28]	否	√	—	—	零样本	27.80	31.50	34.20	34.10
	SemFace ^[29]	是	—	√	√	—	28.80	35.20	34.50	32.90
	Lass ^[30]	是	√	√	—	零样本	—	—	—	—
PTNMT 中 PTM 的引入策略	Imamura 等人 ^[31]	否	—	√	—	—	28.90	—	—	—
	Clichant 等人 ^[32]	否	—	—	—	—	27.90	—	—	—
	XLM-T ^[33]	否	√	√	—	—	—	—	—	—
	Rothe 等人 ^[34]	否	—	—	—	—	31.70	34.20	—	—
	DeltaLM ^[35]	否	√	√	—	—	—	—	—	—
	mBART-50 ^[36]	是	√	√	—	—	—	—	—	—
	CPT ^[37]	是	—	√	—	零样本	—	—	—	—
	AB-Net ^[38]	否	—	√	—	—	30.60	33.57	—	39.21

续表

类别	模型	PTM 为 端到端	其他 NMT 实验				WMT14/%		WMT16/%	
			多语言	低资源	无监督	其他实验	EN-DE	DE-EN	EN-RO	RO-EN
PTNMT 中 PTM 的引入 策略	Graformer ^[39]	否	✓	✓	—	零样本	—	36.90	—	—
	DENOIS ^[40]	是	✓	✓	✓	—	28.90	39.30	—	—
	Stickland ^[41]	是	✓	✓	—	—	—	—	—	—
	Deep-fusing ^[42]	否	—	—	—	—	31.59	—	—	36.71
	BERT-fused ^[43]	否	—	—	✓	篇章翻译	30.75	—	36.02	39.10
	Park 等人 ^[44]	否	—	✓	—	—	—	—	—	—
	Aspect Augmented ^[45]	否	—	✓	—	—	28.05	—	—	—
	BiBERT ^[46]	否	—	✓	—	—	30.91	34.94	—	—
	CTNMT ^[47]	否	—	—	—	—	30.10	—	—	—
	APT ^[48]	否	—	—	—	—	29.23	34.84	—	—
	MSP ^[49]	否	✓	✓	—	—	21.20	—	—	—
	HICTL ^[50]	否	—	✓	—	零样本	31.74	—	—	39.88
	Zhao 等人 ^[51]	是	—	✓	—	—	—	—	—	—
	CRISS ^[52]	是	✓	✓	✓	—	—	—	38.20	38.50
PTNMT 模型及 数据 分析	Wang 等人 ^[53]	是	—	—	—	—	—	—	38.00	38.10
	Susato 等人 ^[54]	是	✓	✓	—	多模态	—	—	—	—
	Aligned-BERT ^[55]	否	—	—	—	—	—	—	—	—
	Zan 等人 ^[56]	是	—	—	—	—	—	—	—	—
	Han 等人 ^[57]	是	—	—	—	—	—	—	—	—
	Copying -penalty ^[14]	是	—	—	—	—	29.6	—	—	—
	Song 等人 ^[58]	是	—	✓	—	零样本	—	—	—	—
	Richardso 等人 ^[59]	是	—	✓	—	零样本	—	—	—	—
	Huang 等人 ^[60]	否	—	✓	—	—	—	—	—	—
	Liu 等人 ^[61]	是	—	—	—	—	—	—	41.6	—
PTNMT 在特定 场景中的 应用	BRLM ^[67]	否	—	—	—	零样本	—	—	—	—
	M.Lakew 等人 ^[68]	否	✓	✓	✓	零样本	—	—	—	—
	SixT ^[69]	否	✓	✓	—	零样本	—	—	—	—
	SixT+ ^[70]	否	✓	✓	✓	零样本	—	—	—	—
	Guo 等人 ^[71]	否	—	—	—	篇章翻译	—	—	—	—
	Multi-context ^[72]	否	—	✓	—	篇章翻译	29.65	—	—	—
	Context -interactive ^[73]	是	—	—	—	篇章翻译	—	—	—	—
	PGC ^[74]	是	—	—	—	篇章翻译	—	—	—	—

续表

类别	模型	PTM 为 端到端	其他 NMT 实验				WMT14/%		WMT16/%	
			多语言	低资源	无监督	其他实验	EN-DE	DE-EN	EN-RO	RO-EN
PTNMT 在特定 场景中的 应用	DEEP ^[75]	是	—	✓	—	实体翻译	—	—	—	—
	Marie 等人 ^[76]	否	—	—	—	—	—	—	—	—
	Han 等人 ^[77]	否	—	—	✓	零样本	—	—	—	—
	Werff 等人 ^[78]	否	—	—	—	人工与机器翻译	—	—	—	—

3 PTNMT 的优缺点及相关分析

将 PTM 引入 NMT,可以帮助 NMT 模型很好地使用无标签数据,提高泛化能力和鲁棒性,实现通用的机器翻译。但同时,在 NMT 中引入 PTM 也会给 NMT 带来一些消极影响,例如模型参数冗余、模型将输入复制作为输出、生成非目标语言的词汇。上述引入 PTM 后出现的问题并不是 NMT 独有的,在自然语言处理的其他领域甚至序列到序列的任务上都存在类似的情况,接下来将对这些情况进行分析和总结。

3.1 引入 PTM 带来的积极影响

3.1.1 使用无标签数据

基于深度学习发展而来的 NMT 十分地依赖数据,本质上是典型的数据驱动任务。在大规模无标签数据上通过自监督训练的 PTM 可以很好地利用无标签数据并学习到通用表示知识。通过对通用表示知识的迁移,一方面针对同语种,可以利用大规模的单语数据提取的通用表示知识增强模型对该语言的建模能力;另一方面针对不同语种,可以利用多语言数据提取的通用表示知识促进模型对该语言的理解。将 PTM 引入下游 NMT 任务从某些角度来讲增加了 NMT 的数据量,这对无论是高资源还是低资源语言都是有益的,但相对而言,对低资源语言的帮助更大。

3.1.2 缓解数据稀缺影响

对于大部分语言而言,存在数据稀缺的问题,这会导致很难充分地训练 NMT 模型。数据稀缺问题导致机器翻译具有一些特殊的场景,例如无监督翻译、低资源翻译、零资源翻译等。在 NMT 中引入 PTM,可以充分利用大规模无标签数据以及 PTM 提取的通用表示知识,并提高这类场景下机器翻译的性能。目前,数据驱动的深度学习,其性能的好坏

很大程度上取决于数据规模的大小,PTM 模型可以缓解数据稀缺影响的特性,不仅在机器翻译中得到重视,而且在自然语言处理领域的其他任务^[10-12]以及计算机视觉领域的任务^[102-104]中均得到广泛应用。

3.1.3 提高模型泛化能力和鲁棒性

语义和领域信息 现有的许多工作^[5-9]表明,在大规模数据上训练 PTM 可以提取数据中的语义和领域信息。同时,随着训练数据的规模增大、领域增多,PTM 获得的信息就更加广泛,能够适配的下游任务环境也越多。这类信息在 PTNMT 中表现十分突出,可以提高 PTNMT 模型的泛化能力。有研究表明,与 NMT 模型相比,PTNMT 模型生成的译文更具多样性,且领域适配性更强。PTNMT 生成的译文更符合语境和实际情况,但这种优势可能无法从评价指标 BLEU 值上体现出来。语义和领域信息不仅在下流 NMT 任务中起作用,在自然语言处理领域的其他任务中也起作用,而 PTNMT 提取和使用这类信息的方法能够给其他任务带来启发。

模型参数分布调优 使用参数初始化和适配器等方法将 PTM 引入 NMT,能够使 NMT 模型参数在初始时处于一个较优的状态,这一方面有利于减少模型的训练代价,无须从零开始训练模型;另一方面有利于模型参数的训练,因为后续模型参数更新可以在这个状态的基础上寻找更优的参数,降低了模型训练初期的不稳定影响,提高了模型的鲁棒性。

限制搜索空间 由于神经机器翻译这类序列到序列生成任务的预测空间极大,因此这类任务的训练过程较为困难,而引入 PTM 可以限制搜索空间,大大降低模型训练的难度,同时通过这种方式可以将 PTM 模型提取的通用表示知识迁移到下游任务上。

3.1.4 通用机器翻译模型

通用表示知识迁移 多语言 PTM 可以从大规

模的多语言数据中学习到不同语言的通用表示知识,这种知识能够有效地帮助 NMT 模型完成源语言到目标语言的转换。通过特定的策略,可以使用多语言 PTM 构建多语言机器翻译模型和通用机器翻译模型,相比于使用多语言数据从零开始训练多语言机器翻译模型,这种方式更为简单高效,且受通用表示知识的影响,可以实现更多语种的多语言机器翻译模型。使用通用表示知识甚至可以使 PTNMT 模型对从未在预训练语料库中出现的语言具备翻译能力。

研究人员发现,对多语言 PTM 通用表示知识的迁移能够大大增强模型对低资源语言的理解,改善低资源机器翻译的性能。通用表示知识的迁移对低资源、中资源、高资源机器翻译有着不同程度的影响,研究人员为了更好地利用通用表示知识研究了多种策略。起初,使用通用表示知识在极高资源机器翻译上没有性能的提高^[19],随着研究的深入,研究人员针对该问题进行改进,如 Li 等人^[26]和 Lin 等人^[30]的方法,最终实现了在极高资源机器翻译上的性能提高。

通用表示知识迁移对其他任务的影响 通用表示知识迁移在预训练的语音翻译中应用较为广泛,语音翻译任务将源语言的语音转换成目标语言的文本,该任务不仅涉及单模态的通用表示知识,还涉及跨模态的通用表示知识。为了更好地完成两种知识的迁移,在预训练的语音翻译中研究人员设计了多种策略,包括针对语音的预训练^[64-65]、针对文本的预训练^[95]、针对语音识别的预训练^[83]、针对文本翻译的预训练^[85],以及使用多种 PTM 搭建语音翻译模型^[96-97]的方法。这些策略的核心思想是提取并迁移通用表示知识,其设计思想与 PTNMT 也有相似之处。除语音翻译外,还有研究人员将 PTM 与多模态机器翻译 (Multimodal Machine Translation, MMT) 相结合^[105-108]。MMT 是指,给定图片和描述图片的源语言,通过计算生成目标语言。由于在 MMT 的输入中包含图像,MMT 也需要使用单模态的通用表示知识和跨模态的通用表示知识。

3.2 引入 PTM 带来的消极影响

模型参数冗余 在大规模数据上训练的 PTM 拥有庞大的参数量,涉及的数据范围越广、模型深度越深,参数量就越庞大,这在多语言 PTM 上体现得更为明显。如此庞大的模型参数量,虽然可以为下游任务提供更多的信息,但同时也带来了不好的影

响。一方面,过大的模型会导致常见的硬件设备无法加载,限制了模型的适用范围;另一方面,PTM 的信息并不全是下游 NMT 任务所需要的,例如将多语言 PTM 使用在双语机器翻译任务中时,PTM 中为除这两种语言外的其他语言建模的参数存在冗余的情况。并且,在大规模数据上训练的 PTM 可能会因为预训练数据中带有偏见数据,导致在下游 NMT 任务应用时产生有失社会公平性的输出。在其他自然语言处理任务上也有类似的问题。

模型复制输入作为输出 由于 PTM 的预训练目标是使用自监督任务预测被掩码的词汇,或根据被破坏文本重建原文本,这会鼓励 PTM 将输入复制作为输出的行为。同样,在 PTNMT 任务中会鼓励模型对一些实体名词或专有名词原封不动地直接复制作为输出,而这些实体名词和专有名词在目标语言中存在另一种特有表述,这种现象会降低译文的可读性和可理解性。有一些工作^[14]针对该问题进行研究,但无法完全消除。

生成非目标语言词汇 使用多语言 PTM 搭建的 PTNMT 模型在翻译过程中存在翻译出非目标语言词汇的情况。例如,构建的是中文翻译到英语的模型,而在生成时产生了法语的词汇。虽然这种现象不会对评价指标 BLEU 产生很大的扰动,但就生成译文的可读性来说,是致命性的打击。

4 未来展望

4.1 PTNMT 实际应用

近几年,“大规模 PTM”是一个非常火热的话题,因其在许多自然语言处理任务上强大的性能表现引起了众多研究人员的关注。机器翻译是自然语言处理领域中较为困难的任务,研究人员期望引入 PTM 以提升模型性能。当前许多 PTNMT 工作的出发点是探索如何利用 PTM 更好地增强 NMT,尚不存在统一的范式,PTNMT 的发展还处于上升阶段。

虽然目前 PTNMT 没有像其他自然语言处理任务一样在引入 PTM 后性能大幅提升,但依然产生了很多可直接用于商用服务的模型,例如 T5 和 GPT-3^[109]等。另外,PTNMT 也获得了更多研究人员的关注,在 2022 年举办的国际机器翻译比赛 (Workshop on Machine Translation, WMT) 中, WMT 主办方允许使用 PTM,这也是 PTNMT 发展

趋势的体现。

当前阶段的 PTNMT 相关探索和研究更多的集中在学术界,距 PTNMT 真正被投入工业界使用还有一些差距。其中一个重要的原因是,工业界使用的 NMT 系统中所涉及的语言大多为高资源语言。高资源语言的数据量足以训练出性能强大的 NMT 模型,再加之当前的方法不能充分利用 PTM 所提供的通用表示知识,且没有统一的范式,这导致了 PTNMT 的商用价值大大下降。但同时,这也为 PTNMT 的未来发展指明了道路。

4.2 PTNMT 前沿问题

虽然 PTNMT 已经取得了巨大的成功,但当前 PTNMT 依然受到很多的限制。

第一,PTM 的预训练任务与下游 NMT 任务存在天然的差距,虽然已有很多工作缓解了该问题带来的负面影响,但是训练目标的差异依然对 PTNMT 造成了困扰,优化 PTM 和 NMT 的目标差异非常具有研究意义。

第二,当前已有部分研究人员在构建 PTM 时使用各种数据资源,包括无监督的单语语料库和监督的双语语料库,已取得了良好的效果。但在 PTNMT 领域中这类方法还较少,该类方法的潜力还有待挖掘。

第三,与 PTM 在自然语言处理的其他领域展现出的强大性能提升不同,PTNMT 在译文质量上没有取得突破性的改进,并且很少有 PTNMT 方法在性能上优于反向翻译等数据增强方法。因此未来需要更深入地分析 PTM 为何不能大幅增强 NMT 的性能,如何改进才能发挥 PTM 的全部能力。

第四,为了在大规模数据上捕获丰富的语义信息,PTM 的参数数量通常较为庞大,这就导致 PTM 模型的参数量比常见的 NMT 模型要大得多,如何有效利用 PTM 仍需进一步研究。

第五,目前存在许多的方法将 PTM 引入 NMT,但这些方法都存在缺点,不能准确地利用 PTM 的信息,且不同研究人员有不同的思路,没有一种较为统一的范式,因此研究如何将 PTM 引入 NMT 的统一范式也极为重要。

5 结语

本文整理了 PTNMT 相关的工作,并从 PTNMT 中 PTM 的预训练策略、PTNMT 中 PTM 的引入策

略、PTNMT 模型行为分析、PTNMT 在特定环境中的应用四个方面系统性地对 PTNMT 工作进行了分类、分析和总结。随后又介绍了在 NMT 中引入 PTM 带来的优缺点,并给出为其他任务带来的启示。最后对 PTNMT 的未来发展进行展望。PTNMT 还有许多问题亟待解决,PTM 在其他领域展现出来的巨大潜力也使 PTNMT 这个研究方向具有很大的实用价值。

参考文献

- [1] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning, 2017: 1243-1252.
- [2] JORDAN M I. Serial order: A parallel distributed processing approach[M]. Advances in Psychology, North-Holland, 1997, 121: 471-495.
- [3] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems, 2017: 6000-6010.
- [4] 肖桐,朱靖波.机器翻译:基础与模型[M].北京:电子工业出版社,2021:373-374.
- [5] KENTON J D M W C, TOUTANOVA L K. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of NAACL-HLT, 2019: 4171-4186.
- [6] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining[C]//Proceedings of the Advances in Neural Information Processing Systems, 2019.
- [7] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI Blog, 2019, 1(8): 9-17.
- [8] SONG K, TAN X, QIN T, et al. Mass: Masked sequence to sequence pre-training for language generation [J]. arXiv preprint arXiv: 1905.02450, 2019.
- [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] LI J, TANG T, ZHAO W X, et al. Pretrained language models for text generation: A survey[C]//Proceedings of the 13th International Joint Conference on Artificial Intelligence, 2021.
- [11] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-

- 1897.
- [12] YUAN S, ZHAO H, ZHAO S, et al. A roadmap for big model[J]. arXiv preprint arXiv: 2203.14101, 2022.
- [13] WANG M, LI L. Pre-training methods for neural machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, 2021: 21-25.
- [14] Liu X, Wang L, Wong D F, et al. On the copying behaviors of pre-training for neural machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 4265-4275.
- [15] REN S, WU Y, LIU S, et al. Explicit cross-lingual pre-training for unsupervised machine translation[J]. arXiv preprint arXiv: 1909.00180, 2019.
- [16] MAO Z, CHU C, KUROHASHI S. Linguistically driven multi-task pre-training for low-resource neural machine translation[J]. Transactions on Asian and Low-Resource Language Information Processing, 2022, 21(4): 1-29.
- [17] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.
- [18] LUO F, WANG W, LIU J, et al. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation[J]. arXiv preprint arXiv: 2010.16046, 2020.
- [19] 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.
- [20] DABRE R, SHROTRIYA H, KUNCHUKUTTAN A, et al. Indicbart: A pre-trained model for natural language generation of indic languages[J]. arXiv preprint arXiv: 2109.02903, 2021.
- [21] YANG J, MA S, ZHANG D, et al. Alternating language modeling for cross-lingual pre-training[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9386-9393.
- [22] ZHANG W, LI X, YANG Y, et al. Pre-training on mixed data for low-resource neural machine translation[J]. Information, 2021, 12(3): 133-139.
- [23] YANG Z, HU B, HAN A, et al. Csp: Code-switching pre-training for neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2624-2636.
- [24] LIN Z, PAN X, WANG M, et al. Pre-training multilingual neural machine translation by leveraging alignment information[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 2649-2663.
- [25] PAN X, WANG M, WU L, et al. Contrastive learning for many-to-many multilingual neural machine translation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 244-258.
- [26] LI P, LI L, ZHANG M, et al. Universal conditional masked language pre-training for neural machine translation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 6379-6391.
- [27] FAN A, BHOSALE S, SCHWENK H, et al. Beyond English-centric multilingual machine translation [J]. Journal of Machine Learning Research, 2021, 22(107): 1-48.
- [28] WANG S, TU Z, TAN Z, et al. Language models are good translators[J]. arXiv preprint arXiv: 2106.13627, 2021.
- [29] REN S, ZHOU L, LIU S, et al. SemFace: Pre-training encoder and decoder with a semantic interface for neural machine translation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 4518-4527.
- [30] LIN Z, WU L, WANG M, et al. Learning language specific sub-network for multilingual machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 293-305.
- [31] IMAMURA K, SUMITA E. Recycling a pre-trained BERT encoder for neural machine translation[C]//Proceedings of the 3rd Workshop on Neural Generation and Translation, 2019: 23-31.
- [32] CLINCHANT S, JUNG K W, NIKOULINA V. On the use of BERT for neural machine translation[C]//Proceedings of the 3rd Workshop on Neural Generation and Translation, 2019: 108-117.
- [33] MA S, YANG J, HUANG H, et al. XLM-T: Scaling up multilingual machine translation with pre-trained cross-lingual transformer encoders[J]. arXiv preprint arXiv: 2012.15547, 2020.
- [34] ROTHE S, NARAYAN S, SEVERYN A. Leveraging pre-trained checkpoints for sequence generation tasks[J]. Transactions of the Association for Compu-

- tational Linguistics, 2020, 8: 264-280.
- [35] MA S, DONG L, HUANG S, et al. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders [J]. arXiv preprint arXiv: 2106.13736, 2021.
- [36] TANG Y, TRAN C, LI X, et al. Multilingual translation with extensible multilingual pretraining and finetuning[J]. arXiv preprint arXiv: 2008.00401, 2020.
- [37] LIU Z, WINATA G I, FUNG P. Continual mixed-language pre-training for extremely low-resource neural machine translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2706-2718.
- [38] GUO J, ZHANG Z, XU L, et al. Incorporating BERT into parallel sequence decoding with adapters [C]//Proceedings of the Advances in Neural Information Processing Systems, 2020: 10843-10854.
- [39] SUN Z, WANG M, LI L. Multilingual translation via grafting pre-trained language models [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 2735-2747.
- [40] ÜSTÜN A, BÉRARD A, BESACIER L, et al. Multilingual unsupervised neural machine translation with denoising adapters [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 6650-6662.
- [41] STICKLAND A C, LI X, GHAZVININEJAD M. Recipes for adapting pre-trained monolingual and multilingual models to machine translation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 3440-3453.
- [42] WENG R, YU H, LUO W, et al. Deep fusing pre-trained models into neural machine translation[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence, 2022.
- [43] ZHU J, XIA Y, WU L, et al. Incorporating BERT into neural machine translation [J]. arXiv preprint arXiv: 2002.06823, 2020.
- [44] PARK J, ZHAO H. Enhancing language generation with effective checkpoints of pre-trained language model[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2686-2694.
- [45] 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, 2021: 2772-2783.
- [46] XU H, VAN D B, MURRAY K. BERT, mBERT, or BiBERT?: A study on contextualized embeddings for neural machine translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 6663-6675.
- [47] YANG J, WANG M, ZHOU H, et al. Towards making the most of BERT in neural machine translation[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9378-9385.
- [48] WENG R, YU H, HUANG S, et al. Acquiring knowledge from pre-trained model to neural machine translation[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9266-9273.
- [49] TAN Z, ZHANG X, WANG S, et al. MSP: Multistage prompting for making pre-trained language models better translators [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 6131-6142.
- [50] WEI X, WENG R, HU Y, et al. On learning universal representations across languages [J]. arXiv preprint arXiv: 2007.15960, 2020.
- [51] ZHAO M, WU H, NIU D, et al. Reinforced curriculum learning on pre-trained neural machine translation models[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9652-9659.
- [52] TRAN C, TANG Y, LI X, et al. Cross-lingual retrieval for iterative self-supervised training[C]//Proceedings of the Advances in Neural Information Processing Systems, 2020: 2207-2219.
- [53] WANG W, JIAO W, HAO Y, et al. Understanding and improving sequence-to-sequence pretraining for neural machine translation [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2591-2600.
- [54] SUSANTO R H, WANG D, YADAV S, et al. Rakuten's participation in WAT: Examining the effectiveness of pre-trained models for multilingual and multimodal machine translation[C]//Proceedings of the 8th Workshop on Asian Translation, 2021: 96-105.
- [55] VÁZQUEZ R, CELIKKANAT H, CREUTZ M, et al. On the differences between BERT and MT encoder spaces and how to address them in translation tasks [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 337-347.
- [56] ZAN C, DING L, SHEN L, et al. On the complementarity between pre-training and random-initializa-

- tion for resource-rich machine translation[C]//Proceedings of the 29th International Conference on Computational Linguistics, 2022: 5029-5034.
- [57] HAN L, EROFEEV G, SOROKINA I, et al. Examining large pre-trained language models for machine translation: What you don't know about it[J]. arXiv preprint arXiv: 2209.07417, 2022.
- [58] SONG H, DABRE R, MAO Z, et al. Pre-training via leveraging assisting languages and data selection for neural machine translation[J]. arXiv preprint arXiv: 2001.08353, 2020.
- [59] RICHARDSON A, WILES J. A systematic study reveals unexpected interactions in pre-trained neural machine translation [C]//Proceedings of the 13th Language Resources and Evaluation Conference, 2022: 1437-1443.
- [60] HUANG D, WANG K, ZHANG Y. A comparison between pre-training and large-scale back-translation for neural machine translation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1718-1732.
- [61] LIU X, WANG L, WONG D F, et al. On the complementarity between pre-training and back-translation for neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 2900-2907.
- [62] SCHNEIDER S, BAEVSKI A, COLLOBERT R, et al. wav2vec: Unsupervised pre-training for speech recognition [C]//Proceedings of the Interspeech, 2019: 3465-3469.
- [63] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//Proceedings of the Advances in Neural Information Processing Systems, 2020: 12449-12460.
- [64] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.
- [65] CHEN S, WANG C, CHEN Z, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.
- [66] CHUNG Y A, ZHANG Y, HAN W, et al. W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training [C]//Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 2021: 244-250.
- [67] JI B, ZHANG Z, DUAN X, et al. Cross-lingual pre-training based transfer for zero-shot neural machine translation[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 115-122.
- [68] LAKEW S M, NEGRI M, TURCHI M. Zero-shot neural machine translation with self-learning cycle[C]//Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages, 2021: 96-113.
- [69] 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 Conference on Empirical Methods in Natural Language Processing, 2021: 15-26.
- [70] CHEN G, MA S, CHEN Y, et al. Towards making the most of cross-lingual transfer for zero-shot neural machine translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 142-157.
- [71] GUO Z, LE NGUYEN M. Document-level neural machine translation using BERT as context encoder [C]//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, 2020: 101-107.
- [72] DONATO D, YU L, DYER C. Diverse pretrained context encodings improve document translation [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 1299-1311.
- [73] YANG P, ZHANG P, CHEN B, et al. Context-interactive pre-training for document machine translation[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2021: 3589-3595.
- [74] CHEN L, LI J, GONG Z, et al. Improving context-aware neural machine translation with source-side monolingual documents[C]//Proceedings of the 13th International Joint Conference on Artificial Intelligence, 2021: 3794-3800.
- [75] HU J, HAYASHI H, CHO K, et al. DEEP: Denoising entity pre-training for neural machine translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 1753-1766.
- [76] MARIE B, FUJITA A. Synthesizing monolingual data for neural machine translation[J]. arXiv preprint arXiv: 2101.12462, 2021.
- [77] HAN J M, BABUSCHKIN I, EDWARDS H, et al.

- Unsupervised neural machine translation with generative language models only[J]. arXiv preprint arXiv: 2110.05448, 2021.
- [78] VAN DER WERFF T, VAN NOORD R, TORAL A. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context[C]//Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, 2022: 161-170.
- [79] 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.
- [80] BANSAL S, KAMPER H, LIVESCU K, et al. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, 2019: 58-68.
- [81] BÉRARD A, BESACIER L, KOCABIYIKOGLU A C, et al. End-to-end automatic speech translation of audiobooks[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2018: 6224-6228.
- [82] LIU Y, ZHANG J, XIONG H, et al. Synchronous speech recognition and speech-to-text translation with interactive decoding[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 8417-8424.
- [83] WANG C, WU Y, LIU S, et al. Curriculum pre-training for end-to-end speech translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3728-3738.
- [84] ALINEJAD A, SARKAR A. Effectively pretraining a speech translation decoder with machine translation data[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 8014-8020.
- [85] DONG Q, WANG M, ZHOU H, et al. Consecutive decoding for speech-to-text translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(14): 12738-12748.
- [86] LIU Y, XIONG H, HE Z, et al. End-to-end speech translation with knowledge distillation[J]. arXiv preprint arXiv: 1904.08075, 2019.
- [87] STOIAN M C, BANSAL S, GOLDWATER S. Analyzing ASR pretraining for low-resource speech-to-text translation[C]//Proceedings of the ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 7909-7913.
- [88] WANG C, WU Y, LIU S, et al. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 9161-9168.
- [89] WANG C, WU A, PINO J, et al. Large-scale self-and semi-supervised learning for speech translation[J]. arXiv preprint arXiv: 2104.06678, 2021.
- [90] WU A, WANG C, PINO J, et al. Self-supervised representations improve end-to-end speech translation[C]//Proceedings of the Interspeech, 2020: 1491-1495.
- [91] CHEN J, MA M, ZHENG R, et al. Mam: Masked acoustic modeling for end-to-end speech-to-text translation[J]. arXiv preprint arXiv: 2101.11445, 2020.
- [92] NGUYEN H, BOUGARES F, TOMASHENKO N, et al. Investigating self-supervised pre-training for end-to-end speech translation[C]//Proceedings of the Interspeech, 2020.
- [93] CHUNG Y A, GLASS J. Generative pre-training for speech with autoregressive predictive coding[C]//Proceedings of the ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 3497-3501.
- [94] LE H, PINO J, WANG C, et al. Lightweight adapter tuning for multilingual speech translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 817-824.
- [95] DONG Q, YE R, WANG M, et al. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021, 35(14): 12749-12759.
- [96] ZHENG R, CHEN J, MA M, et al. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation[C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 12736-12746.
- [97] YE R, WANG M, LI L. End-to-end speech translation via cross-modal progressive training[J]. arXiv preprint arXiv: 2104.10380, 2021.
- [98] XU C, HU B, LI Y, et al. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 2619-2630.
- [99] BAPNA A, CHUNG Y, WU N, et al. SLAM: A u-

- nified encoder for speech and language modeling via speech-text joint pre-training[J]. arXiv preprint arXiv: 2110.10329, 2021.
- [100] AO J, WANG R, ZHOU L, et al. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 5723-5738.
- [101] LI X, WANG C, TANG Y, et al. Multilingual speech translation from efficient finetuning of pre-trained models[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 827-838.
- [102] ZHAI X, KOLESNIKOV A, HOULSBY N, et al. Scaling vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12104-12113.
- [103] RIQUELME C, PUIGSERVER J, MUSTAFA B, et al. Scaling vision with sparse mixture of experts[J]. Advances in Neural Information Processing Systems, 2021, 34: 8583-8595.
- [104] LIU Z, HU H, LIN Y, et al. Swin transformer v2: Scaling up capacity and resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12009-12019.
- [105] CAGLAYAN O, KUYU M, AMAC M S, et al. Cross-lingual visual pre-training for multimodal machine translation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 1317-1324.
- [106] YAWEI K, FAN K. Probing multi-modal machine translation with pre-trained language model[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 3689-3699.
- [107] SONG Y, CHEN S, JIN Q, et al. Product-oriented machine translation with cross-modal cross-lingual pre-training[C]//Proceedings of the 29th ACM International Conference on Multimedia, 2021: 2843-2852.
- [108] HIRASAWA T, KANEKO M, IMANKULOVA A, et al. Pre-trained word embedding and language model improve multimodal machine translation: A case study in multi30K[J]. IEEE Access, 2022, 10: 67653-67668.
- [109] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.



曹智泉(1999—),硕士研究生,主要研究领域为自然语言处理、预训练模型。

E-mail: caozhiquan0519@foxmail.com



肖桐(1982—),通信作者,博士,教授,主要研究领域为自然语言处理。

E-mail: xiaotong@mail.neu.edu.cn



穆永誉(1997—),博士研究生,主要研究领域为自然语言处理、预训练模型。

E-mail: lixiaoyumu9@gmail.com