计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900005

融合特定语言适配模块的多语言神经机器翻译

刘俊鹏[1]　苏劲松[2]　黄德根[1]
1 大连理工大学计算机科学与技术学院　辽宁 大连 116024
2 厦门大学信息学院　福建 厦门 361005
(liujunpeng_nlp@mail.dlut.edu.cn)

摘　要　多语言神经机器翻译利用单一的编码器-解码器模型对多种语言之间的翻译同时进行建模。多语言神经机器翻译不仅能够促进关联语言之间的知识迁移,提高低资源语言的翻译质量,并且能够实现未见语言对之间的翻译。现有多语言神经机器翻译仍然存在语言多样性建模能力不足和未见语言对翻译质量不佳的问题。为此,首先在现有的适配器模型基础上提出变维双语适配器模型,在Transformer模型的每个子层之间加入双语适配器以抽取每个语言对的独特特征,并通过改变适配器隐层维度调整编码器和解码器两端的特定语言表达空间;其次,提出一种共享单语适配器模型,对每种语言的独特特征进行建模。在IWSLT多语言翻译数据集上的实验结果表明,变维双语适配器模型能够显著提升多语言翻译的性能,而单语适配器模型能够在不影响多语言翻译性能的条件下提高未见语言对的翻译质量。

关键词:多语言神经机器翻译;特定语言建模;双语适配器;单语适配器
中图法分类号　TP391

Incorporating Language-specific Adapter into Multilingual Neural Machine Translation

LIU Jun-peng[1],SU Jin-song[2] and HUANG De-gen[1]
1 School of Computer Science and Technology,Dalian University of Technology,Dalian,Liaoning 116024,China
2 School of Informatics,Xiamen University,Xiamen,Fujian 361005,China

Abstract　Multilingual neural machine translation (mNMT) leverages a single encoder-decoder model for translations in multiple language pairs. mNMT can encourage knowledge transfer among related languages,improve low-resource translation and enable zero-shot translation. However,the existing mNMT models are weak in modeling language diversity and perform poor zero-shot translation. To solve the above problems,we first propose a variable dimension bilingual adapter based on the existing adapter architecture. The bilingual adapters are introduced in-between each two Transformer sub-layers to extract language-pair-specific features and the language-pair-specific capacity in the encoder or the decoder can be altered by changing the inner dimension of adapters. We then propose a shared monolingual adapter to model unique features for each language. Experiments on IWSLT dataset show that the proposed model remarkably outperforms the multilingual baseline model and the monolingual adapter can improve the zero-shot translation without deteriorating the performance of multilingual translation.

Keywords　Multilingual neural machine translation,Language-specific modeling,Bilingual adapter,Monolingual adapter

1　引言

近年来,神经机器翻译(Neural Machine Translation,NMT)[1]在众多任务上超越了传统的统计机器翻译方法,成为机器翻译领域的主要方法。现有的NMT模型大多只针对单个语言对的翻译任务进行建模,要实现多种语言之间的翻译常常需要训练多个NMT模型,因此如何简单快速地实现多语言翻译成为当前机器翻译领域的前沿问题。Johnson等[2]通过在源语言句子起始位置添加目标语言标签的方式实现了基于单个编码器-解码器模型的多语言翻译。这种方法不仅能够降低多语言机器翻译模型的训练和部署成本,促进不同语言之间的知识迁移,还能够实现未见语言对之间的翻译(Zero-shot Translation),因而成为多语言神经机器翻译(Multilingual Neural Machine Translation,mNMT)的新范式。

然而多语言神经机器翻译仍然面临一些挑战。一方面,多种语言共享一个编码器-解码器模型,使得多语言翻译模型存在容量瓶颈(capacity bottleneck)问题,导致多语言翻译性

到稿日期:2021-09-01　返修日期:2021-10-19
基金项目:国家重点研发计划(2020AAA0108004)
This work was supported by the National Key Research and Development Program of China(2020AAA0108004).
通信作者:黄德根(huangdg@dlut.edu.cn)

能下降；另一方面，目标语言标签提供的有限语言信息往往不足以对语言的多样性进行建模。因此，在多语言翻译的相关研究中，针对特定语言的信息进行建模是一项重要的研究内容。Bapna 等[3]提出了一个轻量级神经网络模块——面向特定语言对的双语适配器(Bilingual Adapter)，能够在预训练的多语言翻译模型的基础上通过微调的方式显著提升多语言翻译的质量。但是，随着语言数量的增加，该方法在微调过程中引入的参数数量和时间代价也随之增加，并且面向特定语言对的 Bilingual Adapter 模型无法用于处理 zero-shot 翻译问题。

针对上述问题，本文对现有的 Bilingual Adapter 模型进行了改进。首先，在 Transformer 的每个子层之间均添加 Adapter 模块，最大限度地抽取特定语言对的特征信息，并且在训练阶段将 Adapter 模块中的参数与 Transformer 模型中的参数一同初始化训练，以减少微调过程产生的时间代价。其次，在保持模型引入的特定语言容量(language-specific capa-city)总量不变的前提下，对编码器和解码器两端的特定语言容量进行调整，从而得到最佳建模效果。最后，在 Bilingual Adapter 模型的基础上提出一种共享单语适配器(Monolingual Adapter)模型，在不影响多语言翻译性能的条件下，使 Adapter 模型能够应用于 zero-shot 翻译任务。实验结果表明，相比基线系统，本文提出的方法能够显著提升多语言神经机器翻译和 zero-shot 翻译的性能。

2 相关工作

2.1 特定语言信息建模

为了提高多语言神经机器翻译模型对语言多样性的建模能力，降低不同语言之间的干扰，研究者们开始尝试在参数共享的基础上针对特定语言的特征信息进行建模。Wang 等[4]通过引入解码器端语言标签、特定语言位置编码和隐层表示的方式提高“一对多”(One-to-Many)翻译的质量；Sachan 等[5]在 Transformer 模型的基础上对多语言翻译中的参数共享策略进行研究，提出利用部分参数共享的方法来提高不同语系语言之间的翻译质量；Platanios 等[6]利用源语言和目标语言的语言向量(language embedding)通过上下文参数生成器(contextual parameter generator)生成编码器和解码器的参数，并控制参数共享的方式；Tan 等[7]利用语言向量将需要翻译的语言划分成不同的语言簇(language cluster)，通过为每个语言簇中的语言各自训练一个多语言翻译模型的方式促进同一语言簇内的知识迁移，并降低不同语言簇间的干扰；Wang 等[8]利用统一表示器(universal representor)替换原有的编码器和解码器模型，以实现参数共享的最大化，通过引入语言敏感的词嵌入表示、注意力机制和判别器等方式提高模型对不同语言的感知和建模能力；Bapna 等[3]在多语言预训练翻译模型的基础上引入轻量级的 Adapter 模块，通过微调的方式提高每个语言对的翻译质量；Zhang 等[9]通过语言感知归一化网络(Language-Aware Layer Normalization, LALN)将不同语言映射到不同的高斯空间，并利用语言感知线性变换(Language-Aware Linear Transformation, LALT)对不同语言之间的词序关系进行建模；Zhang 等[10]利用门机制训练多语言翻译模型，实现共享或特定语言路径的动态选择。

2.2 Zero-shot 翻译

尽管多语言神经机器翻译模型能够实现 zero-shot 翻译，但其翻译结果中常常存在错译为另一种语言的现象，即翻译脱靶(off-target translation)问题。如表 1 所列，多语言神经机器翻译模型将源语言翻译为英语，而非指定的罗马尼亚语。此外，zero-shot 翻译还存在源语言和生成语言之间的伪关联(spurious correlation)[11]、语言无关表示的成分缺失(missing ingredient)[12]等问题。为此，Currey 等[13]提出借助中间语言来改善未见语言对的翻译质量；Firat 等[14]、Gu 等[11]和 Zhang 等[9]通过利用回译(back-translation)方法构造伪双语数据的方式来减少翻译脱靶现象的发生；Arivazhagan 等[12]通过引入对齐正则项(alignment regularizer)损失训练模型学习语言无关表示，以提升模型的泛化能力；Al-Shedivat 等[15]通过定义一致性概念，利用基于一致协议(consistency agreement)的学习方法鼓励模型用第 3 种语言生成与平行语料等同的翻译信息；Philip 等[16]提出一种面向特定语言的单语适配器模型，在预训练的多语言翻译模型基础上通过微调的方式提升 zero-shot 翻译质量。

表 1 荷兰语→罗马尼亚语 zero-shot 翻译中的翻译脱靶问题示例

Table 1 Example of off-target translation issue with Dutch→Romanian zero-shot translation with a MNMT model

源语言句子	Dus, je kunt het doen
参考译文	Deci, poti să o duci la îndeplinire
Zero-shot 译文	So, you can do it

3 多语言神经机器翻译模型

3.1 Transformer 模型

基于 Transformer 模型的神经机器翻译模型因其良好的翻译性能成为目前主流的翻译架构。Transformer 模型的编码器和解码器均由多个相同的子层堆叠而成。编码器的每个子层包含两个基本模块，分别为多头自注意力模块和前馈神经网络模块。解码器的每个子层除了有掩码多头自注意力模块和前馈神经网络模块之外，在两者之间还有一个跨语言多头注意力模块。每个基本模块内均使用残差网络和层归一化(layer normalization)进行连接。因此，每个基本模块内的网络可以简化地表示为：

$$h^{\text{out}} = LN(h^{\text{in}} + f(h^{\text{in}})) \tag{1}$$

其中，h^{in}和h^{out}分别表示每个基本模块的输入和输出，$LN(\cdot)$表示层归一化，$f(\cdot)$表示多头注意力网络或前馈神经网络。

3.2 多语言神经机器翻译

为了使神经机器翻译模型能够同时处理多个语言对的翻译，Johnson 等[2]提出在每个源语言句子的起始部分添加一个语言标签(language tag)，用于表示该源语言句子所要翻译成的目标语言。给定源语言句子 $X'=\{x_1, x_2, \cdots, x_m\}$及其对应的目标语言标签 $lang$，多语言翻译的源语言输入被改写为 $X=\{lang, x_1, x_2, \cdots, x_m\}$。

多语言神经机器翻译使用同一个编码器和解码器模型对

不同语言之间的翻译进行建模。在训练阶段，模型在所有语言对上对参数进行优化，使得总体损失达到最小。多语言神经机器翻译的训练目标函数表示如下：

$$L(D;\theta)=\sum_{l=1}^{L}\sum_{d=1}^{|D_l|}\sum_{t=1}^{M}\log P(y_t^l\,|\,x^l,y^l<t;\theta) \tag{2}$$

其中，L 表示语言对个数，D 表示多语言翻译训练集，M 表示目标句子长度，θ 表示模型参数。

4 融合特定语言适配器的多语言神经机器翻译模型

为了提升多语言神经机器翻译模型捕捉不同语言之间的差异性特征的能力，本文提出融合特定语言适配器(language-specific adapter)的多语言神经机器翻译模型，其模型结构如图 1 所示。

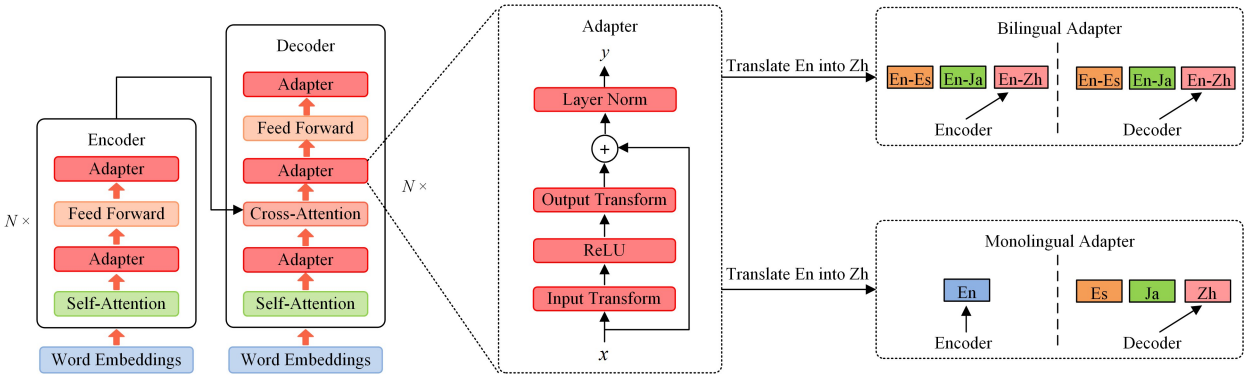

图 1 融合特定语言适配器模块的多语言神经机器翻译模型

Fig. 1 Multilingual neural machine translation model with language-specific adapter

4.1 Adapter 模块结构

Adapter 模型常被应用于多领域多任务学习任务。Bapna 等[3]对 Adapter 模型的结构进行了改进并将其应用于多语言翻译任务。与之相类似，本文使用一个单层的前馈神经网络作为 Adapter 模型的基本模块。为了防止梯度消失和梯度爆炸问题，在 Adapter 模块中加入了残差连接和层归一化网络。如图 1 所示，给定任意 Transformer 子层的输出 h_i，经过 Adapter 模块后的输出 $\tilde{h}_i$ 的计算方式如下：

$$\tilde{h}_i=LN(h_i+g(h_i)) \tag{3}$$

$$g(h_i)=\max(0,h_iW_1+b_1)W_2+b_2 \tag{4}$$

其中，$W_1\in\mathbb{R}^{d\times d_f}$，$W_2\in\mathbb{R}^{d_f\times d}$，$b_1\in\mathbb{R}^{d_f}$ 和 $b_2\in\mathbb{R}^{d}$ 为训练参数，d 表示模型隐层维度，d_f 表示 Adapter 模型中前馈神经网络的隐层维度。

4.2 面向语言对的变维双语适配器模型

多语言翻译任务可以看作一类特殊的多任务学习问题。在同一个翻译模型下，多个翻译任务通过联合训练得到一个“平均”模型，导致其模型参数偏离了双语翻译模型，从而损失了每个翻译任务各自的建模特征。因此，首先研究面向特定翻译任务(语言对)的变维双语适配器模型。在 Transformer 模型的每个子层之间均插入一个 Bilingual Adapter 模块，用于抽取特定语言对的独有特征信息。在训练阶段，所有 Adapter 模块中的参数与原有的多语言翻译模型参数一同初始化训练。由于 Adapter 模块的引入会增加模型参数，为了控制模型的参数总量，在编码器和解码器两端分别设置一组 Adapter 模块参数，且每组参数在编码器或解码器的各个子层之间共享。对于不同的翻译任务，其编码器和解码器所需的特定语言表达空间可能存在差异。通过调整编码器和解码器两端 Adapter 模块中的前馈神经网络的隐层维度 d_f^e 和 d_f^d，可以调整编码器和解码器两端语言特定表达空间的分配。为了保持模型中特定语言表示空间总量恒定，规定 d_f^e，d_f^d 和 d_f 三者之间的关系如式(5)所示：

$$2\,d_f=d_f^e+d_f^d \tag{5}$$

4.3 共享单语适配器模型

尽管 Bilingual Adapter 模型能够抽取特定语言对的特征信息，但由于缺少相应的 Adapter 模块参数，该模型不能应用于未见语言对的翻译任务。针对该问题，本文提出一种共享单语适配器(Shared Monolingual Adapter)模型(以下简称 Monolingual Adapter)。具体地，Monolingual Adapter 模型为每一种语言 l_k 设置一组参数 θ_k，在将语言 l_i 翻译为语言 l_j 时，利用源语言 l_i 和目标语言 l_j 对应的参数 θ_i 和 θ_j，分别初始化编码器和解码器中的 Adapter 模块，用来抽取每种语言的特征信息。这样的模型结构使Mo-nolingual Adapter 模型既能够应用于多语言翻译任务，也能够用于 zero-shot 翻译。Monolingual Adapter 模型的共享性体现在两个方面：一是模型为每种语言只保留一组 Adapter 参数，该参数在编码器和解码器两端共享；二是 Adapter 参数在编码器和解码器的各个子层之间共享。本文提出的 Monolingual Adapter 与 Philip 等[16]的方法存在以下 4 个不同点：一是 Adapter 模块的作用位置不同，所提方法在 Transformer 的任意两个子层之间均插入 Monolingual Adapter 模块以提升模型对语言特有信息的建模能力；二是 Adapter 模块数量不同，本文为 n 种语言设置 n 组 Adapter 参数，而 Philip 等的方法为 $2n$；三是训练方式不同，本文的 Adapter 参数与模型中的其他参数一同初始化训练，而未采用“预训练-微调”的训练范式；四是实验设置不同，本文研究并验证了在没有语言标签作为目标语言指示的条件下，利用 Monolingual Adapter 模块为 zero-shot 任务提供语言信息的可行性。

4.4 Bilingual Adapter 与 Monolingual Adapter 的参数比较

假设共有 n 种语言，对于“一对多”(One-to-Many)或“多

对一”(Many-to-One)翻译任务,Bilingual Adapter需要保存$(n-1)$组*Adapter*模块参数,而*Monolingual Adapter*模型则需要保存n组参数;对于“多对多”翻译任务,Bilingual Adapter模型需要保存$n(n-1)$组参数,而Monolingual Adapter模型只需保存n组参数。

5 实验

5.1 实验数据及参数设置

对于One-to-Many和Many-to-One翻译任务,实验数据采用2011—2018年IWSLT评测任务数据集,其包含英语与其他16种语言的平行语料。对于zero-shot翻译任务,实验数据采用IWSLT-17评测数据集,通过训练英语与其他3种语言的“多对多”翻译模型,实现3种非英语语言之间6个方向的zero-shot翻译。数据集的具体统计信息如表2和表3所列。

表2 IWSLT数据集统计

Table 2 Statistics of IWSLT dataset

Language Pair	Train/($\times10^3$)	Valid	Test
En-Ar	223	887	1 569
En-Cs	114	480	1 511
En-De	196	887	1 565
En-Es	180	887	1 570
En-Fr	219	887	1 664
En-He	184	888	1 568
En-It	181	887	1 529
En-Ja	221	871	1 549
En-Nl	167	887	1 569
En-Pl	176	767	1 564
En-Ro	181	887	1 567
En-Ru	177	887	1 568
En-Sl	17	1 144	1 411
En-Tr	154	887	1 568
En-Vi	129	768	1 342
En-Zh	208	887	1 570

表3 IWSLT-17数据集统计

Table 3 Statistics of IWSLT-17 dataset

Language Pair	Train/($\times10^3$)	Valid	Test
En—It	231.6	929	1 566
En—Nl	237.2	1 003	1 777
En—Ro	220.5	914	1 678
It—Ro	217.5	914	1 643
Nl—Ro	206.9	913	1 680
It—Nl	233.4	1 001	1 669

实验的基线系统采用基于Transformer的多语言翻译模型。对于One-to-Many和Many-to-One任务,按照文献[7]进行了参数设置,具体而言,编码器和解码器的层数L均设置为2,隐层状态和词嵌入维度d均设置为256,前馈神经网络隐层维度d_{ffn}为1 024。而对于zero-shot任务,上述3个参数分别设置为6,512和2 048[8]。所有句子首先使用Moses工具分词,而后利用字节对编码(Byte-Pair-Encoding,BPE)[17]进行子词切分,词汇表大小设置为64 000,源端和目标端句子的最大长度设置为100,每个batch中包含8192个token。在训练过程中使用Adam算法[18]对参数进行优化,dropout比率为0.1,初始学习率为5×10^{-4}。测试阶段使用集束搜索(beam search)算法和长度惩罚(length penalty)机制进行解码,其中,集束宽度为4,长度惩罚系数为0.6。实验结果使用SacreBLEU[19]工具进行评价,并区分字母大小写。

为了验证所提方法在多语言翻译任务上的效果,本文设计了多个基线系统进行对比实验。

(1)Bilingual[1]:为每个语言对各自训练一个神经机器翻译模型。

(2)Multilingual[2]:基于Transformer的多语言翻译模型,对于One-to-Many和Many-to-One任务,语言标签分别表示目标语言和源语言。

(3)Multilingual-w:在Multilingual模型的基础上将模型隐层维度由256增加至288,以保持模型参数量与所提方法相当。

(4)Multilingual-d:在Multilingual模型的基础上将编码器和解码器的层数由2增加至4,以保持模型参数量与所提方法相当。

(5)Adapter[3]:在Transformer每一层的顶端为每个语言对加入一个适配器模块,其隐层维度d_f设置为128。

(6)Language Cluster[7]:首先利用语言向量将所有语言划分为不同的语言簇,并为每个语言簇单独训练一个多语言翻译模型。语言簇的划分结果如表4所列。

表4 语言簇划分结果

Table 4 Language clustering results

Cluster	One-to-Many	Many-to-One
1	Es,Fr,It,Ro	De,Es,Fr,It,Nl,Ro
2	De,Nl	Ar,He
3	Cs,Pl,Ru,Sl	Cs,Pl,Ru,Sl
4	Ja,Tr,Zh	Ja,Tr
5	Ar,He,Vi	Vi
6	—	Zh

(7)LALN+LALT[9]:在多语言翻译基线系统的基础上引入语言敏感的LALN和LALT模块。

5.2 One-to-Many和Many-to-One的实验结果

模型在One-to-Many和Many-to-One任务上的实验结果如表5所列。其中,*BLEU*值表示各个模型在16个语言对上的平均*BLEU*值,*WM*和*WB*分别表示翻译质量优于多语言翻译基线模型(模型2)和双语翻译基线模型(模型1)的语言对数量。由表3的实验结果可以得到以下结论:1)多语言翻译基线模型译文的平均*BLEU*值与双语翻译基线模型相当,但超过一半语言对的翻译质量仍然不如双语翻译模型。2)通过增加模型隐层维度(模型3)和模型深度(模型4)的方式扩大多语言翻译模型容量,可以提高One-to-Many任务的翻译质量,但对于Many-to-One任务无提升效果,甚至出现翻译质量下降的现象。其原因可能在于训练集数量较小,增大模型容量后导致过拟合现象发生,因而造成翻译质量下降。3)Bilingual Adapter模型能够稳定地提升One-to-Many和Many-to-One两个翻译任务的译文质量,并且系统性能优于微调方

法（模型 5）。与模型 3 相比，Bilingual Adapter 模型在使用更少的模型参数的条件下取得了更好的提升效果。而与模型 4 相比，Bilingual Adapter 模型尽管在 One-to-Many 任务上表现略差，但在 Many-to-One 任务上则表现更好。综合而言，在模型参数接近的情况下，Bilingual Adapter 模型能够带来更加稳定的提升效果，同时证明了对特定语言空间建模的必要性。

4）Monolingual Adapter 模型（模型 9）也能够提升多语言神经机器翻译模型的性能，但与 Bilingual Adapter 模型相比，由于使用了更少的模型参数，因此翻译性能略差。在将 Monolingual Adapter 模型的参数量增加至与 Bilingual Adapter 模型（模型 10）接近后，二者的翻译性能也接近。这一结果证明了 Monolingual Adapter 模型的有效性。

表 5 One-to-Many 和 Many-to-One 任务 BLEU 的评测结果

Table 5 Evaluation results of BLEU on One-to-Many and Many-to-One task

ID	模型	L	d	d_f^e	d_f^d	参数量/($\times 10^6$)	One-to-Many			Many-to-One		
							BLEU	WM	WB	BLEU	WM	WB
1	Bilingual	2	256	—	—	981.76	19.01	9	—	22.70	11	—
2	Multilingual	2	256	—	—	61.36	19.25	—	7	22.72	—	5
3	Multilingual-w	2	288	—	—	69.26	20.15	16	12	22.70	7	5
4	Multilingual-d	4	256	—	—	65.05	21.35	16	14	21.73	2	5
5	2+Adapter	2	256	128	128	63.48	19.68	13	9	23.19	15	8
6	2+Language Cluster	2	256	—	—	306.80	20.74	14	15	23.55	11	11
7	2+LALN+LALT	2	256	—	—	62.49	19.90	16	9	23.24	12	7
8	Bilingual Adapter	2	256	256	256	65.58	**20.84**	16	14	**23.83**	15	10
9	Monolingual Adapter(256)	2	256	256	256	63.60	19.79	15	8	23.67	16	9
10	Monolingual Adapter(512)	2	256	256	512	65.70	20.80	16	13	**23.83**	15	11

5.3 Bilingual Adapter 数量与性能的分析

与 Bapna 等只在 Transformer 编码器和解码器的每一层后加入适配器模块的方案（记作“+Top Adapter”）不同，本文方法在 Transformer 模型的每个基本模块之后均加入了适配器模块（记作“+All Adapter”）。为此，对两种方案的性能进行对比分析，实验结果如表 6 所列。由表 6 的结果可知，在翻译性能上，“+All Adapter”模型在 Many-to-One 任务上较“+Top Adapter”模型的 *BLEU* 值提升了 0.31 个百分点，但在 One-to-Many 任务上则降低了 0.18 个百分点。在训练时间方面，尽管两种方案的参数总量相同，但由于“+All Adapter”模型中适配器数量较“+Top Adapter”模型有所增加，因此训练时间也有所增加，约为“+Top Adapter”模型的 1.3 倍。

表 6 Bilingual Adapter 数量与性能关系的分析

Table 6 Analysis on relationship between amount of Bilingual Adapter and performance

系统	参数量/($\times 10^6$)	训练时长	One-to-Many	Many-to-One
Multilingual	61.36	1.0×	19.25	22.72
+Top Adapter	65.58	1.2×	**21.02**	23.52
+All Adapter	65.58	1.5×	20.84	**23.83**

5.4 Bilingual Adapter 模型消融实验

为了研究编码器和解码器对语言特定空间的需求情况，对 Bilingual Adapter 模型进行了消融实验，结果如表 7 所列。

表 7 Bilingual Adapter 模型的消融实验

Table 7 Ablation study of Bilingual Adapter

	One-to-Many	Many-to-One
Multilingual	19.25	22.72
Encoder	20.07	23.61
Decoder	20.26	23.09
Both	**20.84**	**23.83**

由表 7 可知，对于 One-to-Many 和 Many-to-One 任务，在编码器或解码器端加入 Bilingual Adapter 模块后译文结果的 *BLEU* 值较多语言翻译基线系统均获得显著提升，并且在编码器和解码器两端同时加入 Bilingual Adapter 模块时提升效果最佳。对于 One-to-Many 任务，在解码器端加入 Bilingual Adapter 模块后的提升效果略优于编码器，原因在于解码器端需要更多的语言特定空间来对不同的语言进行建模，从而生成不同语言的译文结果；而 Many-to-One 任务则相反，在编码器端加入 Bilingual Adapter 模块后的提升效果优于解码器，这是由于此时解码器端只生成一种目标语言，建模难度相对较小，因此对语言特定空间的需求较少。

5.5 语言特定空间分配对翻译质量的影响

现有的 Bilingual Adapter 模型在编码器和解码器端引入的语言特定空间往往是相同的。然而，由 5.4 节的实验结果可知，对于不同的翻译任务，编码器和解码器端对语言特定表示空间的需求可能不同。为了验证该假设，在保证模型引入语言特定空间总量不变的条件，对编码器或解码器两端特定语言空间分配对翻译质量的影响进行了研究，实验结果如图 2 所示。

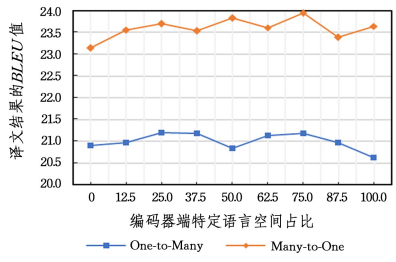

图 2 翻译结果与编码器端特定语言空间占比的关系

Fig. 2 Translation results when varying rate of language-specific capacity in encoder

由图 2 可知，对于 One-to-Many 任务，当编码器和解码器

端的特定语言空间比例为1:3时，Bilingual Adapter模型取得最佳的翻译效果；而Many-to-One任务则相反，当Bilingual Adapter模型获得最佳翻译效果时，编码器和解码器端特定语言空间的比例为3:1。上述实验结果表明，对于不同翻译任务，编码器和解码器端对语言特定表示空间的需求也各不相同。

5.6 zero-shot实验结果

由于Bilingual Adapter模型无法用于zero-shot翻译，因此仅使用Monolingual Adapter模型对zero-shot翻译任务进行改进，实验结果如表8所列。由表8的结果可知，尽管多语言神经机器翻译模型能够实现zero-shot翻译，但是其翻译结果的*BLEU*值较相应的双语翻译模型降低了6.87个百分点。而与基于中间语言的翻译模型(模型2)相比，多语言翻译模型在*BLEU*值指标上也有1.38个百分点的差距。加入Monolingual Adapter模块后，zero-shot的翻译质量较多语言翻译模型(模型3)的*BLEU*值提高了2.38个百分点，并超过了基于中间语言的翻译模型。Gu等[11]的研究表明，由于目标语言标签与编码器语言建立了错误的联系，导致zero-shot任务的翻译质量下降。为了研究Monolingual Adapter模块是否可以作为语言标记来表示源语言所要翻译成的目标语言，在实验中去掉了源语言句首的语言标签(模型5)。由表8可知，在去掉语言标签后，zero-shot的翻译质量进一步提高了0.38个百分点。这表明Monolingual Adapter模型一方面能够替代语言标记用于指示目标语言的种类，另一方面可以减少伪关联现象的发生。尽管Monolingual Adapter模型能够提升zero-shot任务的翻译质量，但是其结果与双语翻译模型相比仍有较大差距，因此zero-shot翻译问题仍有待进一步研究。

表8 IWSLT-17数据集上的未见语言对翻译结果

Table 8 Translation results of zero-shot task on IWSLT-17 dataset

ID	模型	It→Ro	Ro→It	Nl→Ro	Ro→Nl	It→Nl	Nl→It	AVG.
1	Bilingual	18.37	22.12	17.86	19.84	19.84	20.29	19.72
2	Pivot	11.15	14.59	11.41	14.31	19.71	14.19	14.23
3	Multilingual	12.45	12.98	12.75	13.25	12.81	12.83	12.85
4	Monolingual Adapter	**14.51**	16.24	13.63	16.13	15.96	15.03	15.23
5	Monolingual Adapter w/o LE	14.44	**16.61**	**14.07**	**16.54**	**16.63**	**15.35**	**15.61**

结束语 多语言神经机器翻译由于无法有效地对语言的多样性进行建模而导致其翻译质量往往不如双语翻译模型。针对上述问题，本文以Adapter模块为基础，通过研究Adapter模块的使用方式、语言特定表达空间分配等问题，提出可变维度的Bilingual Adapter模型，在保持特定语言表达空间不变的前提下进一步提升了多语言神经机器翻译的性能。在此基础上，针对多语言神经机器翻译在zero-shot翻译任务上表现不佳的问题，本文提出Monolingual Adapter模型，使得模型在保持One-to-Many和Many-to-One两个任务的翻译质量不受影响的同时提升未见语言对的翻译质量。

目前，本文Adapter模型在部分语言对上的翻译质量与双语模型仍有差距。在未来的工作中，我们将在大规模多语言翻译任务上研究更加灵活的语言特定空间建模方法，进一步提升多语言神经机器翻译的性能。

参 考 文 献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[2] JOHNSON M, SCHUSTER M, LE Q V, et al. Google's Multilingual Neural Machine Translation System: Enabling zero-shot Translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.

[3] BAPNA A, ARIVAZHAGAN N, FIRAT O. Simple, Scalable Adaptation for Neural Machine Translation[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). Hong Kong: ACL, 2019: 1538-1548.

[4] WANG Y N, ZHANG J J, ZHAI F F, et al. Three Strategies to Improve One-to-Many Multilingual Translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 2955-2960.

[5] SACHAN D C, NEUBIG G. Parameter Sharing Methods for Multilingual Self-Attentional Translation Models[C]// Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels: ACL, 2018: 261-271.

[6] PLATANIOS E A, SACHAN M, NEUBIG G, et al. Contextual Parameter Generation for Universal Neural Machine Translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 425-435.

[7] TAN X, CHEN J L, HE D, et al. Multilingual Neural Machine Translation with Language Clustering[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). Hong Kong: ACL, 2019: 963-973.

[8] WANG Y N, ZHOU L, ZHANG J J, et al. A Compact and Language-Sensitive Multilingual Translation Method[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 1213-1223.

[9] ZHANG B, WILLIAMS P, TITOV I, et al. Improving Massively

Multilingual Neural Machine Translation and Zero-Shot Translation[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: ACL. 2020: 1628-1639.

[10] ZHANG B,BAPNA A,SENNRICH R,et al. Share or not? Learning to Schedule Language-specific Capacity for Multilingual Translation[C] // International Conference on Learning Representations. 2021.

[11] GU J T,WANG Y,CHO K,et al. Improved zero-shot Neural Machine Translation via Ignoring Spurious Correlations[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL,2019:1258-1268.

[12] ARIVAZHAGAN N,BAPNA A,FIRAT O,et al. The Missing Ingredient in zero-shot Neural Machine Translation[J] arXiv: 1903.07091.

[13] CURREY A,HEAFIELD K. Zero-resource Neural Machine Translation with Monolingual Pivot Data[C]// Proceedings of the 3rd Workshop in Neural Generation and Translation. Hong Kong: ACL,2019:99-107.

[14] FIRAT O,SANKARAN B,AL-ONAIZAN Y,et al. Zero-resource Translation with Multi-lingual Neural Machine Translation[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016: 268-277.

[15] AL-SHEDIVAT M,PARIKH A. Consistency by Agreement in zero-shot Neural Machine Translation[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL,2019:1184-1197.

[16] PHILIP J,BÉRARD A,GALLÉ M,et al. Monolingual Adapters for Zero-shot Neural Machine Translation[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online: ACL,2020:4465-4470.

[17] SENNRICH R,HADDOW B,BIRCH A. Neural Machine Translation of Rare Words with Subword Units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL,2016:1715-1725.

[18] KING D P,BA J. Adam: A Method for Stochastic Optimization [J]. arXiv:1412.6980.

[19] POST M. A Call for Clarity in Reporting BLEU Scores [C]// Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels: ACL,2019:186-191.

LIU Jun-peng, born in 1992, postgraduate. His main research interests include machine translation and so on.

HUANG De-gen, born in 1965, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine translation and neural language processing.

(责任编辑:李亚辉)