文章编号：1003-0077(2020)07-0001-18

神经机器翻译前沿综述

冯 洋[1,2]，邵晨泽[1,2]

（1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院大学，北京 100049）

摘 要：机器翻译是指通过计算机将源语言句子翻译到与之语义等价的目标语言句子的过程，是自然语言处理领域的一个重要研究方向。神经机器翻译仅需使用神经网络就能实现从源语言到目标语言的端到端翻译，目前已成为机器翻译研究的主流方向。该文选取了近期神经机器翻译的几个主要研究领域，包括同声传译、多模态机器翻译、非自回归模型、篇章翻译、领域自适应、多语言翻译和模型训练，并对这些领域的前沿研究进展做简要介绍。

关键词：神经机器翻译；模型训练；同声传译；多模态机器翻译；非自回归机器翻译；篇章翻译；领域自适应；多语言翻译

中图分类号：TP391 **文献标识码**：A

Frontiers in Neural Machine Translation: A Literature Review

FENG Yang[1,2], SHAO Chenze[1,2]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Machine translation is a task which translates a source language into a target language of the equivalent meaning via a computer, which has become an important research direction in the field of natural language processing. Neural machine translation models, as the main stream in the reasearch community, can perform end-to-end translation from source language to target language. In this paper, we select several main research directions of neural machine translation, including model training, simultaneous translation, multi-modal translation, non-autoregressive translation, document-level translation, domain adaptation, multilingual translation, and briefly introduce the research progresses in these directions.

Keywords: neural machine translation; model training; simultaneous translation; multi-modal translation; non-autoregressive translation; document-level translation; domain adaptation; multilingual translation

1 神经机器翻译

机器翻译是指通过计算机将源语言句子翻译到与之语义等价的目标语言句子的过程，是自然语言处理领域的一个重要研究方向。1949 年，Warren Weaver 在《翻译》中提出了使用机器进行翻译的思想，自此引发了该方向的研究热潮。机器翻译主要可以分为三种方法：基于规则的机器翻译、基于统计的机器翻译和基于神经网络的机器翻译。最初，基于规则的方法是机器翻译研究的主流，这种方法对语法结构规范的句子有较好的翻译效果，但其也有规则编写复杂、难以处理非规范语言现象的缺点。20 世纪 90 年代初，IBM 的 Peter Brown 等人发表了两篇重要论文[1-2]，正式提出基于噪声信道模型的统计机器翻译模型。进入 21 世纪，深度学习等机器学习方法逐渐成熟，并开始被应用于自然语言处理领域。2013 年，Kalchbrenner 和 Blunsom 提出利用神经网络进行机器翻译[3]，随后一两年内，Sutskever[4]、Cho[5-6]、Bahdanau[7] 等人提出了基于编码器—解码器

收稿日期：2020-01-06 **定稿日期**：2020-01-22

基金项目：国家重点研发计划政府间国际科技创新合作重点专项(2017YFE0192900)

结构的神经机器翻译模型,标志着机器翻译进入深度学习的时代。2016年,Junczys Dowmunt等人[8]在30多个语言对上对神经机器翻译和统计机器翻译进行对比,神经机器翻译在27个任务上超过了基于短语的统计机器翻译,这展现了神经机器翻译的强大能力。

尽管神经机器翻译已经表现出比统计机器翻译更加优异的翻译效果,但其仍具有巨大的发展潜力。2016年,Wu等人[9]公布了谷歌的神经机器翻译模型,该模型通过在层之间引入残差连接解决了深度模型梯度消失的问题,将模型层数堆叠到了8层,使机器翻译的水平提升到了一个新的台阶。随后,facebook的Gehring等人[10]提出了基于卷积神经网络的编码器—解码器模型,在准确度上超越了谷歌的模型,并大幅提升了翻译速度。2017年,Vaswani等人[11]提出了基于注意力机制的Transformer模型,在模型的训练速度和翻译质量上都取得了大幅提升。2018年,Hassan等人[12]将多种算法结合,并在翻译评价中引入人工评测,首次宣布模型在新闻领域的翻译上达到了人类水平。2019年以来,又有许多神经机器翻译的模型结构被人们提出[13-19],它们在实验中展现出了高于基准Transformer模型的翻译质量。除此之外,反向翻译[20-21]、数据筛选[22-23]、预训练[24]等技术也对翻译效果的提升有显著作用。

尽管神经机器翻译在标准数据集上达到了相当高的翻译质量,但在实际应用中,仍有许多问题需要解决。神经机器翻译模型存在训练和测试时行为不一致的问题,该问题被称为"曝光偏差",引发了研究者的广泛关注;在同声传译的场景下,为了降低翻译的延迟,模型需要在输入语句不完整的情况下输出译文,使用户能低延迟地收到高质量的翻译结果;除文本外,有时也存在图像、视频等其他模态的数据可供翻译模型使用,翻译系统可以融入这些信息以进一步地提高翻译质量;为了提升翻译速度,非自回归模型对翻译概率独立建模,因此能够并行解码出整句译文,但也会出现严重的漏译、过译现象;在对篇章文本进行翻译时,为了保证译文的一致性,模型在翻译时也需同时考虑上下文的信息;存在低资源的领域内数据和高资源的领域外数据时,为了提升模型在领域内的翻译质量,需要对领域外数据也进行合理利用;当需要在多个语言之间进行翻译时,训练多语言翻译模型可以大幅减少所需的翻译模型数目,同时提升低资源语言的翻译质量。

在接下来的内容中,我们将首先介绍神经机器翻译主流框架的演变,随后针对本文上面提到的问题,对神经机器翻译中同声传译、多模态机器翻译、非自回归模型、篇章翻译、领域自适应、多语言翻译和模型训练等方向的前沿研究进展做简要介绍。

2 主流框架

2.1 基于循环神经网络的神经机器翻译模型

基于循环神经网络和编码器—解码器结构的神经机器翻译模型[4-7]在很长一段时间内都是神经机器翻译的主流模型。其中,Bahdanau等人[7]在编码器—解码器框架的基础上,提出了RNNSearch模型,该模型引入了注意力机制,使得生成每个目标端词语时,解码器可以将"注意力"集中到源端的几个相关的词语上,并从中获取有用的信息,从而获得更好的翻译表现。注意力机制使得翻译模型能够更好地处理长距离的依赖关系,解决了在循环神经网络中信息在长距离的传输中容易被丢失、遗忘的问题。RNNSearch模型被研究者广泛地用作基线模型,其注意力机制如图1所示,下面我们对其做具体介绍。

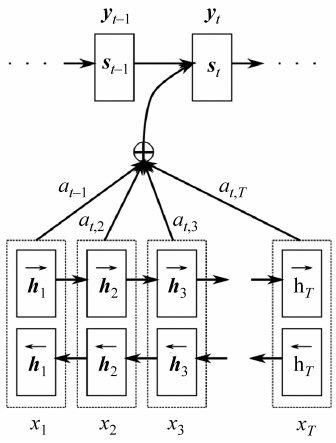

图1 循环神经网络中的注意力机制

(图片引自文献[7])

编码器 使用双向门控循环单元(GRU)[5]对源语句进行编码,以使每个位置的编码同时包含前、后文本的历史信息。双向GRU由前向GRU和后向GRU组成,前向GRU从左向右读取源语句并计算一系列前向隐状态($\overrightarrow{\boldsymbol{h}}_1,\cdots,\overrightarrow{\boldsymbol{h}}_n$),反向GRU从右向左扫描源语句,计算一系列后向隐状态($\overleftarrow{\boldsymbol{h}}_1,\cdots,\overleftarrow{\boldsymbol{h}}_n$)。假设输入序列的词嵌入为($\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$),则编码器如式(1)所示。

$$\overrightarrow{\boldsymbol{h}}_i=\overrightarrow{\mathrm{GRU}}(\boldsymbol{x}_i,\overrightarrow{\boldsymbol{h}}_{i-1}),\quad \overleftarrow{\boldsymbol{h}}_i=\overleftarrow{\mathrm{GRU}}(\boldsymbol{x}_i,\overleftarrow{\boldsymbol{h}}_{i+1}),$$
$$\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}}_i,\overleftarrow{\boldsymbol{h}}_i]。 \tag{1}$$

解码器　解码器是一个前向 GRU，逐词预测译文 $\boldsymbol{y}$。生成译文第 j 个词 y_j 的概率为：

$$P(y_j \mid \boldsymbol{y}_{<j}, \boldsymbol{x}, \theta) = \mathrm{softmax}(\boldsymbol{t}_{j-1}, \boldsymbol{c}_j, \boldsymbol{s}_j) \quad (2)$$

其中，$\boldsymbol{t}_{j-1}$ 是词 y_{j-1} 的词嵌入，$\boldsymbol{s}_j$ 是解码器在第 j 步时的隐状态，$\boldsymbol{c}_j$ 为第 j 步的注意力向量。状态 $\boldsymbol{s}_j$ 计算如式(3)所示。

$$\boldsymbol{s}_j = \mathrm{GRU}(\boldsymbol{t}_{j-1}, \boldsymbol{s}_{j-1}, \boldsymbol{c}_j)。 \quad (3)$$

注意力机制被用于提取与当前步预测高度相关的源端信息，防止源端信息在长距离的解码中被遗忘一部分。在第 j 步的解码中，与位置 i 的源端信息的相关度如式(4)所示。

$$e_{ij} = \boldsymbol{v}_a^{\mathrm{T}} \tanh(\boldsymbol{W}_a \boldsymbol{s}_{j-1} + \boldsymbol{U}_a \boldsymbol{h}_i) \quad (4)$$

注意力向量为源端信息按相关度的加权和如式(5)所示。

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{i'=1}^{n} \exp(e_{i'j})}, \quad \boldsymbol{c}_j = \sum_{i=1}^{T} \alpha_{ij} \boldsymbol{h}_i \quad (5)$$

2.2　基于卷积神经网络的神经翻译模型

在基于循环神经网络的神经机器翻译模型展现出强大的能力后，研究者们也开始探索基于卷积神经网络的神经翻译模型。Meng 等人[25]用卷积神经网络来编码源端语言，并将其集成到统计机器翻译模型中。Gehring 等人[26]将神经网络翻译模型的源端编码器替换成了基于卷积神经网络的结构，随后 Gehring 等人[10]提出了完全基于卷积神经网络的机器翻译模型。下面，我们对文献[10]中的模型做简要介绍。

编码器　卷积神经网络在处理输入序列中的一个片段时，并不知道这个片段在句子中的具体位置。因此，在词嵌入中加入位置编码可以使模型获得更丰富的信息，增强模型的表示能力。假设 $\boldsymbol{w}$ 为词嵌入，$\boldsymbol{p}$ 为位置嵌入，则模型输入为 $\boldsymbol{w}$ 与 $\boldsymbol{p}$ 之和如式(6)所示。

$$\boldsymbol{w} = (w_1, \cdots, w_m), \quad \boldsymbol{p} = (p_1, \cdots, p_m),$$
$$\boldsymbol{e} = (w_1 + p_1, \cdots, w_m + p_m) \quad (6)$$

模型的编码器就是多个卷积模块的叠加，通过卷积操作对输入序列进行编码。令输入窗口大小为 k，模型维数为 d，则卷积模块的输入 $\boldsymbol{X} \in R^{kd}$，卷积核的大小为 $\boldsymbol{W} \in R^{2d \times kd}$，将输入编码为 $2d$ 长度的向量，并通过 GLU 非线性变换[27]将其变换为 d 维向量。随后，卷积模块的输入通过残差连接[28]与 GLU 输出相连，得到卷积模块的输出。

解码器　解码器与编码器结构基本相同，也是由多个卷积模块叠加而成，这里只列出其不同点。在卷积模块的编码后，解码器的第 l 层对编码器的第 u 层做注意力，首先对解码器的隐变量 $\boldsymbol{h}_i^l$ 做变换，如式(7)所示。

$$d_i^l = \boldsymbol{W}_d^l \boldsymbol{h}_i^l + \boldsymbol{b}_d^l + \boldsymbol{g}_i \quad (7)$$

其中，g_i 为目标端第 i 个词的嵌入，W、b 为线性变换的参数。求出对源端第 u 层的注意力，如式(8)所示。

$$a_i^j = \frac{\exp(\boldsymbol{d}_i^l \times \boldsymbol{z}_j^u)}{\sum_{j=1}^{m} \exp(\boldsymbol{d}_i^l \times \boldsymbol{z}_j^u)}, \quad (8)$$

其中，z_j^u 为编码器第 u 层第 j 步的输出。用式(8)所得的注意力权重更新解码器的隐状态，如式(9)所示。

$$c_i^l = \sum_{j=1}^{m} a_{ij}^l (z_j^u + e_j) \quad (9)$$

在更新隐状态时，用到的不仅有编码器输出 z_j^u，还有源端词嵌入 e_j。所得结果 c_i^l 将作为下一个卷积模块的输入。图 2 展示了基于卷积神经网络的编码器-解码器结构及其注意力机制。

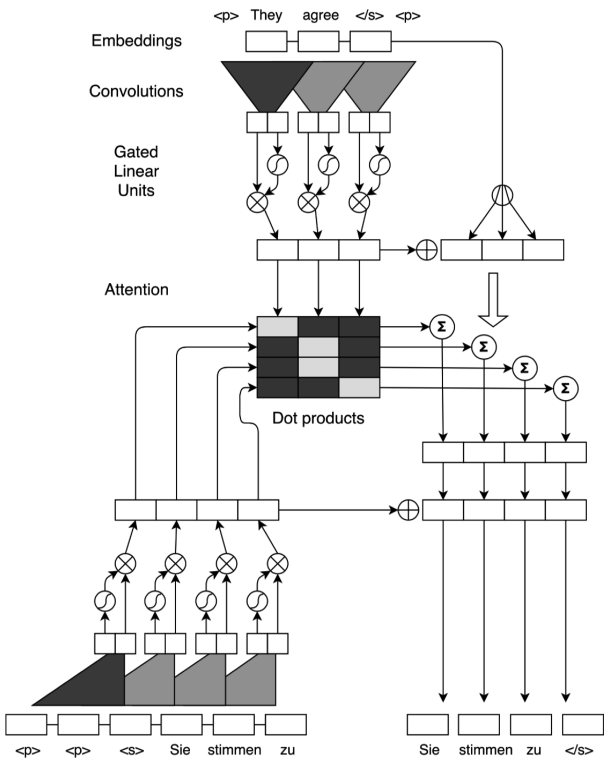

图 2　卷积神经网络中的注意力机制

（图片引自文献[10]）

2.3　Transformer 模型

2017 年，Vaswani 等人提出了完全基于注意力机制的 Transformer 模型[11]，该模型创新性地使用

了自注意力机制来对序列进行编码，其编码器和解码器均由注意力模块和前向神经网络构成。Transformer模型具有高度并行化的模型结构，因此在训练速度上远超循环神经网络，且在翻译质量上也有大幅提升。近期，Transformer已成为神经机器翻译研究中的主流模型，且在自然语言处理的其他领域中也有广泛应用。Transformer的模型结构如图3所示，下面我们对其做具体介绍。

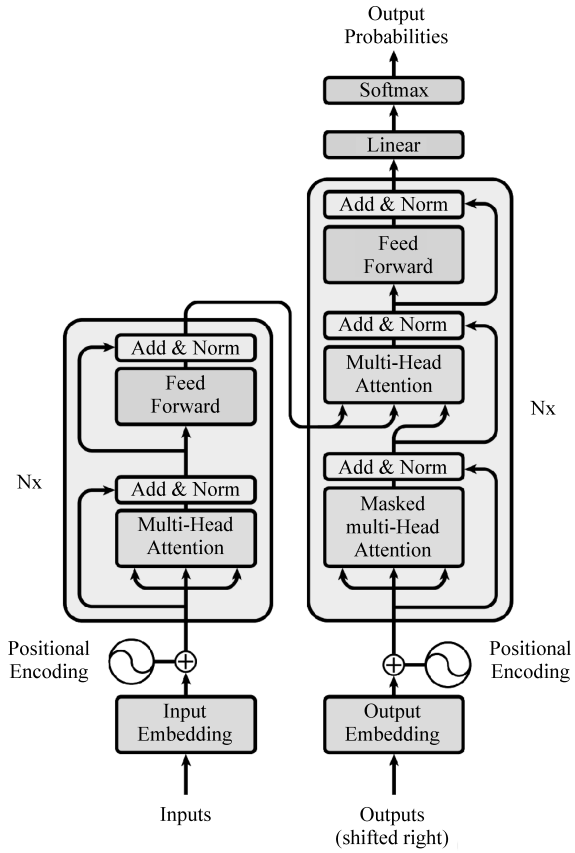

图3 Transformer模型

（图片引自文献[11]）

编码器 由于Transformer模型中没有对序列顺序的显式利用，因此模型需要对词进行位置编码，以表示序列中不同词的位置关系，如式(10)所示。

$$PE(pos,2i)=\sin(pos/10\ 000^{2i/d_{model}}),$$

$$PE(pos,2i+1)=\cos(pos/10\ 000^{2i/d_{model}}) \quad (10)$$

其中，pos表示词在句子中的位置，i表示维数。将位置编码和原本的词嵌入相加后，输入到模型的编码器中。编码器由n个结构相同的层组成，每层主要包含两个模块：注意力和前馈神经网络。注意力模块以点乘注意力为基础，对输入的请求$\boldsymbol{Q}$、键$\boldsymbol{K}$和值$\boldsymbol{V}$做如下操作，如式(11)所示。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (11)$$

其中，d_k为键$\boldsymbol{K}$的维度。在编码器中，使用的是自注意力模块，即将模型输入同时作为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，让输入序列对自身进行注意力计算。除此之外，作者还提出了多头的机制，即在计算注意力时，把输入平均分成多个部分，每个部分独立地计算注意力，最后把得到的注意力结果做拼接，作为最终结果。在计算完自注意力后，用如下的前馈神经网络对输入做变换，如式(12)所示。

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2 \quad (12)$$

除上述自注意力模块和前馈神经网络模块，残差连接[28]、层归一化机制[29]也被应用到了Transformer模型中。

解码器 Transformer的解码器与编码器结构基本相同，这里只列出其不同点。在解码时，译文是单向生成的，前面的解码步骤不应看到后面步骤的翻译结果。因此，在训练时，解码器到自注意力模块里会引入一个单向的mask矩阵，使从前往后的注意力结果被固定为0。另外，在解码器的自注意力模块与前馈神经网络模块之间，还有一个编码器-解码器注意力模块，将解码器的“注意力”集中到源端相关的词语上。其中，解码器的输入为式(11)中$\boldsymbol{Q}$，编码器的输出同时为$\boldsymbol{K}$、$\boldsymbol{V}$。

2.4 性能比较

对于这三种基本模型，在WMT 2014英语到德语的翻译任务上，按照文献[9-11]的配置，采用BLEU-4作为评价指标，性能如表1所示。

表1 主流框架性能比较

模型	En->De
RNNSearch[9]	25.82
ConS2S[10]	25.16
Transformer(base)[11]	27.3
Transformer(big)[11]	28.4

3 同声传译

在同声传译的场景下，为了降低翻译的延迟，模型需要在输入语句不完整的情况下输出译文，使用户能在低延迟内收到高质量的翻译结果。翻译质量和延迟一般是互斥的，在翻译前等待的时间越长，得到的源端信息越完整，翻译质量通常就越高，但延迟也就越高。因此，模型需要在翻译质量和时间延迟

上进行权衡，找出质量和延迟之间的平衡点。在翻译质量上，一般还是用机器翻译中常用的 BLEU 值[30]作为评价指标。在延迟的评价上，研究者提出了 average proportion（AP）[31]、consecutive wait（CW）[32]、average lagging（AL）[33]、differentiable average lagging(DAL)[34]等指标，目前还没有达成一致。在评价同传翻译模型的效果时，通常会结合翻译质量和时间延迟两者，画出延迟-BLEU 值的曲线图，认为曲线整体在上方的模型效果更好。

在同声传译中，一个主要的难点就是让模型决定是否在当前位置输出翻译。如果翻译过早，系统读取的输入不够，会导致翻译质量较低。如果翻译过晚，虽然能获得足够的信息，但是会造成翻译的延迟较高。因此，读写策略的制定也成了研究者重点关注的内容。Cho 和 Esipova[31]首次提出了基于神经机器翻译的同声传译，并提出了一种基于模型翻译概率的变化来制定读写策略的方法。Gu 等人[32]则是通过将翻译质量 BLEU 和延迟指标 AP、CW 设定为奖赏值，运用强化学习算法来让模型自动地学习读写策略。相比于文献[31]的静态策略，文献[32]提出的方法能通过调节翻译质量和延迟奖赏的比值来调节翻译延迟，因此可以适应不同场景下的同传需求。上面两种方法共同的缺点是：它们均是直接使用整句的翻译模型来翻译不完整的输入句，导致训练和测试的模型行为不匹配。文献[33,35]提出了基于固定延迟的读写策略，使得译文总是落后原文固定数目个单词。这种固定的读写策略可以在给定的延迟下训练和测试模型，但也使得模型无法针对特定输入适当地加快或放慢速度。Zheng 等人[36-37]提出了两种对读写策略的监督学习方法，文献[36]中提出的方法使得模型能够尽量输出延迟范围内的动作序列，文献[37]从整句翻译模型的概率变化中分析出动作序列，以此为监督来训练读写策略模块。Arivazhagan 等人[34]使用单调的硬注意力机制来得到读写策略，并在训练中使用软注意力来进行模拟，使得模型的翻译部分和读写策略能同时进行训练。为了使得延迟可控，文献[34]引入了可导的延迟指标 DAL，令模型在训练时也对延迟进行优化。Ma 等人[38]随后将该单调注意力方法扩展到了 Transformer 模型的多头注意力机制中。

除读写策略外，Alinejad 等人[39]在文献[32]的基础上加入了 Predict 操作，使用语言模型来预测当前输入的下一个单词，使模型能获得更完整的输入。束搜索解码算法通常难以应用在同传场景中，Zheng 等人[40]对束搜索算法进行改进，使之能在有限步内进行推测，得到更准确的翻译结果。

4　多模态机器翻译

除文本外，有时也存在图像、视频等其他模态的信息可供使用，多模态翻译系统同时将源端文本和其他模态的信息作为模型输入，翻译系统在其他模态信息的辅助下进行翻译，如图 4 所示。WMT 在 2016 到 2018 年连续三次将多模态机器翻译作为共享任务[41-43]，这项任务也受到了研究者的广泛关注[44-45]。

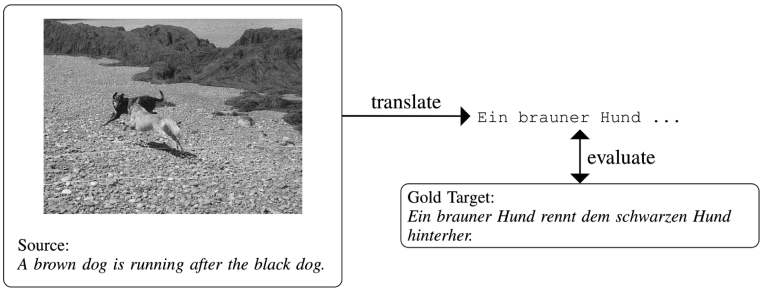

图 4　多模态翻译任务，以图像和源端文本为输入，输出目标端译文
（图片引自文献[41]）

在多模态机器翻译中，一般是以图像为额外的模态信息，辅助翻译系统进行翻译。在对图像进行表示时，通常是用从大型视觉数据集上训练的卷积神经网络(CNN)来提取图像的多种深层特征。为了在翻译模型中应用图像信息，早期的工作通常将图像信息作为输入语句的一部分[46-47]，或用其对编码器、解码器的状态做初始化[46,48-49]。随后，人们开始尝试通过注意力机制来挖掘视觉信息的作用。Caglayan 等人[50]提出了多模态的注意力机制，让模型在解码时同时对文本表示和图像表示做注意力，更充分地对图像信息进行利用。Calixto 等人[51]提出了一种类似的双注意力解码器，通过对文本和图像分别去求注意力，并通

过门结构来控制图像注意力的权重。Delbrouck 等人[52]对多种注意力机制进行了实验对比。Libovicky 等人[53]对文本表示和图像表示独立去求注意力，并应用层次注意力来整合两种注意力结果。Zhou 等人[54]使用一种视觉注意力机制，建立了视觉和文本的联合语义嵌入。除注意力外，Ive 等人[55]提出了一种二阶段的解码方法，在第一阶段仅使用普通的翻译模型，在第二阶段用视觉信息来改善翻译结果。Toyama 等人[56]在多模态翻译中通过变分自编码器的方式引入了隐变量，隐变量中包含文本和图像的信息。在测试时，模型只通过文本预测隐变量，不需要用到图像信息。Calixto 等人[57]对上述方法进行改进，要求模型能从隐变量中重构出图像。Yang 等人[58]同时训练正向和反向的翻译模型，并通过图像信息来对翻译模型做正则化。

5 非自回归模型

目前主流的神经机器翻译模型为自回归模型，每一步的译文单词的生成都依赖于之前的翻译结果，因此模型只能逐词生成译文，翻译速度较慢。如图 5 所示，Gu 等人[59]提出的非自回归神经机器翻译模型(NAT)对目标词的生成进行独立的建模，因此能够并行解码出整句译文，显著提升了模型的翻译速度。然而，非自回归模型在翻译质量上与自回归模型有较大差距，主要表现为模型在长句上的翻译效果较差，译文中通常包含较多的重复词和漏译错误，如图 6 所示。

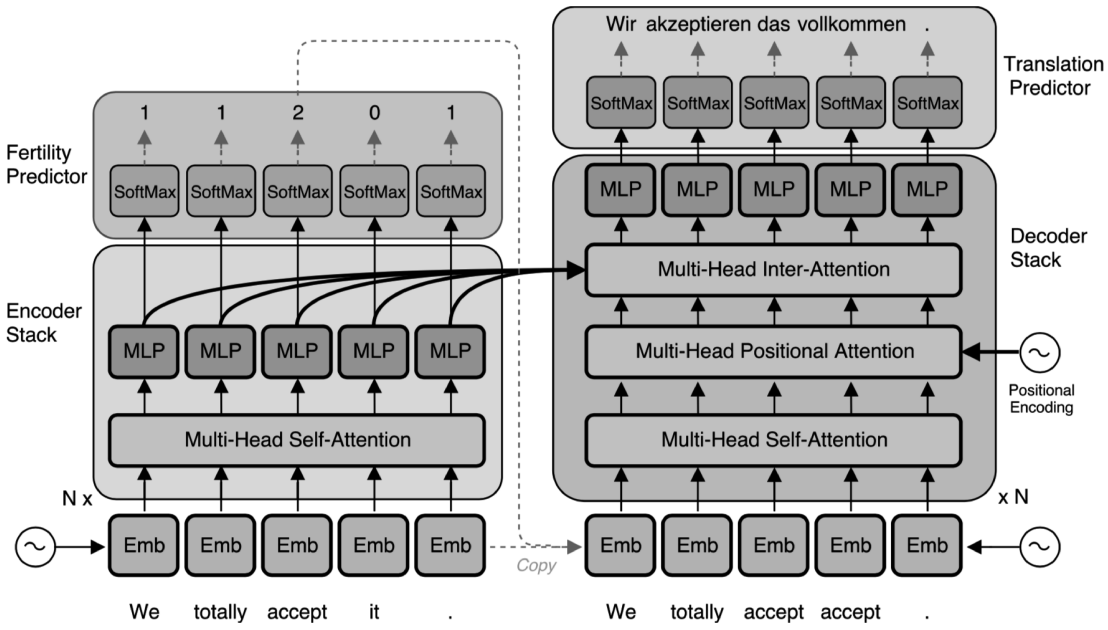

图 5 Gu 等人提出的非自回归翻译模型

(图片引自文献[59])

Src	und noch tragischer ist , dass es Oxford war ···
Ref	even more tragic is that it was Oxford ···
NAT	and more more more more that it was Oxford ···
AR	and , more tragic , Oxford was ···

Src—原文；Ref—参考译文；NAT—非自回归模型；AR—自回归模型

图 6 NAT 译文中的重复词和漏译错误

(图片引自文献[60])

要提升非自回归模型的翻译质量，最简单有效的方法是进行序列级的知识蒸馏[61]。Gu 等人[59]指出，在给定原文时，目标端参考译文的概率分布具有“多峰性”，即一句原文可能对应多句意思相近的译文。由于无法进行 teacher forcing 训练，译文的“多峰性”会对模型的训练造成很大干扰。用序列级知识蒸馏的方法，将平行语料目标端替换为自回归模型的输出，可以有效地缓解译文的“多峰性”问题。Zhou 等人[62]对序列级知识蒸馏做了进一步的探究，指出知识蒸馏能降低目标端文本的“复杂度”，从而降低模型的学习难度。利用自回归模型辅助非自回归模型训练的方法还包括模仿学习[63]、课程学习[64]、提示方法[65]等。Wei 等人[63]提出模仿学习的方法，让非自回归模型学习自回归模型输出的动作序列，从而提高模型的翻译效果。Guo 等人[64]让自回归模型与非自回归模型共享参数，通过课程学习使模型从自回归过渡到非自回归。Li 等人[65]在

训练中利用自回归模型的提示，让模型学习自回归模型隐变量和注意力的分布，以此增强模型能力。

除此之外，提升非自回归模型翻译质量的方法可以大体上分为三类：引入隐变量、改进训练目标和迭代式解码。Gu 等人[59]首先提出引入隐变量来减小译文的不确定性，以此克服译文的“多峰性”问题。他们用产出率(fertility)来表示原文单词对应到译文单词的数目，并将其作为隐变量。给定产出率时，译文空间会受到很大限制，从而缓解了译文的“多峰性”问题。随后，离散隐变量被引入到了非自回归模型中[66-67]，并取得了显著的效果提升。文献[68-69]将变分自编码器[70]引入到了非自回归模型中，对隐变量的先验和后验分布进行建模，并在训练时从后验分布采样隐变量，以此来减小译文的不确定性。此外，Bao 等人[71]将位置信息建模为隐变量，Ran 等人[72]将重排序信息建模为隐变量，都取得了明显的效果提升。引入隐变量也有一定缺陷，即对隐变量的建模会降低非自回归模型的翻译速度。除隐变量外，对训练目标的改进也能有效地提升翻译质量，且一般不会影响翻译速度。Wang 等人[73]直接在训练目标中加入正则化项，以此来抑制模型的重复翻译和漏译错误。Sun 等人[74]用 CRF 对翻译概率进行建模，并用对应的损失函数来优化模型。Libovicky 等人[75]将 CTC 损失应用在非自回归模型中，使得模型可以产生空单词，摆脱对译文长度预测的依赖。Shao 等人[60]在模型中融入序列信息，改进强化学习算法来进行序列级训练，用对译文“多峰性”不敏感的序列级指标来训练模型。Shao 等人[76]随后针对非自回归模型提出了 Bag-of-Ngrams 训练目标，该训练目标具有可导、计算方便、与翻译质量相关性高等优点，能够稳定、高效地训练非自回归模型，相对强化学习方法，翻译效果有大幅提升。对非自回归模型的另一个改进思路为迭代式的解码，即通过多轮迭代，在牺牲一定翻译速度的前提下提升翻译质量。Lee 等人[77]首次提出了迭代式的解码方法，并给出了自适应地控制迭代轮数的方法。Ghazvininejad 等人[78]提出了基于 Mask-Predict 的迭代解码方法，在每一轮的解码结果中掩盖掉模型不确定的部分，并在下一轮对其进行预测。Gu 等人[79]提出 Levenshtein Transformer 模型，对译文进行多轮的删除和插入操作，直至译文不再改变。目前，迭代式的解码技术已能让非自回归模型达到接近自回归模型的翻译效果，并仍保持着数倍于自回归模型的解码速度。

6　篇章翻译

尽管神经机器翻译模型能在单句翻译上达到很好的效果，但真实的文本并不是由孤立、无关的句子组成的。篇章中的不同句子会互相联系，句子的翻译结果也可能会受到其上下文的影响。如果模型忽略了句子与其上下文的联系，仅从句子级别进行翻译，就有可能产生在句子层面正确、但在篇章层面上不合适的翻译结果[80]。Jean 等人[81]首次在神经机器翻译中验证了引入上下文信息的有效性。他们对基于 RNN 的句子级神经机器翻译模型进行扩展，引入了额外的编码器来对源端句子的上下文进行建模，使模型在解码时能同时考虑源端句子和上下文信息。之后的篇章翻译模型也基本沿袭了这个思路，即用额外的模块对上下文进行编码，将上下文信息融入句级的模型结构中。Bawden 等人[82]进一步地对这种多编码器模型进行探索，测试了模型在不同的注意力机制下的翻译质量。Wang 等人[83]用一个两层的层级 RNN 模型对上下文信息进行建模，其中底层的 RNN 对句子进行编码，顶层的 RNN 将所有上下文句子的信息总结为一个向量。在解码器端，他们设计了不同的注意力机制来让模型同时也对上下文向量做注意力。Voita 等人[84]提出了基于 Transformer 的篇章翻译模型，模型的编码器分为上下文编码器和原文编码器，两者共享前 $n-1$ 层的参数，在最后一层将信息通过门结构融合后输出。Zhang 等人[85]也在 Transformer 模型的基础上引入了上下文编码器，从而构建篇章级的神经机器翻译模型。与文献[84]不同的是，文献[85]没有将上下文信息与原文进行融合，而是在编码器、解码器中都加入了针对上下文的注意力模块。Miculicich 等人[86]将层次注意力机制运用到了 Transformer 中，对多句上下文进行建模。Maruf 等人[87]用稀疏注意力对层次注意力机制做改进，使得模型能集中注意力在上下文的相关句子上。Yang 等人[88]提出了一种新型的胶囊网络架构[89-90]，用以建模上下文中每个句子内部词与词以及词与待翻译句子之间的语义关系，使模型能更好地建模上下文单词间的联系。

除上述的引入额外编码器的方法外，研究者还提出了许多基于其他方法的篇章翻译模型。Tu 等人[91]用连续的缓存来存储双语上下文的隐层表示，使模型能通过对缓存做注意力来找出相关的上下文信息。Kuang 等人[92]用动态高速缓存来存储之前

句子的翻译结果和一些语义相关的译文，并在每个解码步骤中对缓存中的单词进行评分，并用门结构将其与模型输出评分组合。Maruf 等人[93]用外部记忆来存储文档级的信息，使模型能通过对外部记忆中的句子做粗略的注意力来捕获全局的上下文信息。Xiong 等人[94]通过模型的两轮解码来鼓励译文生成的一致性。在第一轮的解码中，模型仅用 Transformer 模型做句级的解码。在第二轮中，模型在教师模型奖赏值[95]的指导下完善初始翻译。Voita 等人[80]针对译文的一致性问题建立了一系列的测试集，并在这些测试集上验证了文中提出的二轮解码模型的有效性。随后，Voita 等人[96]指出了文档级平行语料资源稀缺的问题，并提出使用句级平行语料和文档级单语语料训练文档级翻译模型的方法。

7 领域自适应

训练语料的规模会对机器翻译模型的翻译质量造成很大影响，神经机器翻译模型对其尤为敏感。需要进行特定领域上的翻译时，由于领域内的语料数目十分有限，在领域内语料上训练的模型通常都表现不佳，在领域外语料上训练的翻译模型则会遇到领域不匹配的问题，翻译效果也不好。通过大规模领域外语料改善领域内翻译性能的方法被称为领域自适应。由于领域内数据稀缺和领域不匹配的问题在现实世界中很常见，领域自适应具有很高的应用价值，是神经机器翻译中备受关注的一个研究方向。

按任务形式分，可将神经机器翻译中的领域自适应分为有监督和无监督两类。其中，在有监督的场景下，我们能获取到大规模的领域外平行语料和小规模的领域内平行语料；在无监督的场景下，我们仅能获取到领域外平行语料和领域内的单语语料。按所用方法分，可将神经机器翻译中的领域自适应方法分为基于数据的方法和基于模型的方法。基于数据的方法主要通过对领域外语料做筛选来扩充领域内的数据规模，基于模型的方法则通过改进模型的结构、训练方法、解码方法等来提升领域内的翻译性能。我们下面按基于数据和基于模型的分类来介绍领域自适应方面的工作。

早在 2010 年，Moore 和 Lewis[97]就提出了通过数据筛选来增强领域内语言模型性能的方法。他们用领域内外的语言模型分别对领域外的数据进行评分，并将评分差值较小的数据筛选出来，认为它们比较接近目标领域。随后，Axelrod 等人[98]将这种方法扩展到了统计机器翻译中，在源语言和目标语言上均计算领域内、外语言模型的评分差，综合两者来筛选出接近目标领域的句对。在神经机器翻译时代，这种数据筛选的方法仍然非常有效，被广泛应用于各个翻译系统中[5-7]。Wang 等人[99]提出了用句嵌入来做数据筛选的方法，他们利用神经机器翻译模型学得源端句子的向量表示，并从领域外语料中筛选出在向量表示上与领域内较为接近的部分。Van der Wees 等人[100]指出，静态的数据筛选方法不适合神经机器翻译模型的训练，并提出了逐步筛选语料的动态数据筛选方法。对于无监督的领域自适应，Hu 等人[101]用词汇归纳的方法来抽取领域内的词汇，然后对领域内的目标端单语语料做基于词的反向翻译，以此构建领域内的伪平行语料。Chu 等人[102]、Imankulova 等人[103]将其他语言对的语料也看作是领域外数据，把多语言翻译和领域自适应结合了起来。

基于模型的领域自适应方法能按训练方法、解码方法、模型结构进行分类。在训练方法上，领域内的微调[104-106]是最简单的改进方法。这种方法先在全部数据上训练得到领域外的翻译模型，再将训练数据限制在领域内，对模型进行微调。然而，这种方法会使得微调后的模型在领域外的翻译质量大幅下降。Dakwale 和 Monz[107]对这种微调方法做了改进，通过领域外模型的知识蒸馏使得模型也能保持较好的领域外翻译性能。Chu 等人[108]提出了混合微调方法，在微调时同时使用领域内外的数据，但增大了领域内数据的比重。Barone 等人[109]在微调时加入正则化项，以减小模型在领域内的过拟合现象。Wang 等人[110]在全部语料上进行训练，但给不同句子赋予不同的学习率权重，权重根据文献[98]中的数据筛选指标来设定。Chen 等人[111]也使用了类似的思路，不同的是，他们训练了一个领域分类器，根据分类器的评分来设置数据的权重。Yan 等人[112]将权重的设置精确到了词级别上，用领域内、外语言模型的评分差来为每个词设置学习权重。Vilar[113]对领域外模型的隐状态赋予权重，将模型调整到领域内的翻译中。近期，Zhang 等人[114]将课程学习应用到了领域自适应中，通过课程的设计使模型能逐步从领域外过渡到领域内。Zeng 等人[115]提出了迭代式的训练方法，让领域内和领域外的模型迭代地基于知识蒸馏的方法学习对方的知识。在解码方法

上，Gulcehre 等人[116]将领域内的目标端语言模型融入模型解码端，浅层融入的方法将解码器和语言模型预测的概率做加权求和，用深层融入的方法拼接解码器和语言模型的隐状态后预测概率。Dou 等人[117]在文献[116]的基础上也引入了领域外的目标端语言模型，并将这个方法应用在无监督的领域自适应上。Freitag 和 Al-Onaizan[104]在解码时对微调后的模型和领域外的模型做了模型融合。Khayrallah 等人[118]提出了一种基于堆栈的词片解码算法，其中词片由统计机器翻译模型生成。在模型结构上，除文献[116-117]中提出的深层融合模型外，Britz 等人[119]将不同领域的数据混合来进行训练，并同时引入判别器来判断数据所属的领域。Kobus 等人[120]在输入的词嵌入上拼接上一个表示领域的嵌入，使模型能够区分输入数据所属的领域。Thompson 等人[121]和 Wuebker 等人[122]指出，领域外模型的大部分参数均可被固定，仅需对小部分参数在领域内做微调。Gu 等人[123]将模型结构分为共享部分和非共享部分，领域内、外的数据共同使用共享部分的参数来抽取领域无关的特征，并使用自身领域的参数来抽取领域相关的特征。他们也引入了对抗的训练方式，希望判别器无法从共享层中判断当前数据所属领域。

8　多语言翻译

神经机器翻译模型最初只能在两种语言间进行翻译，无法进行多语言的互译[4-7]。这样的模型有两个缺陷：①如果我们需要实现 n 种语言之间的互相翻译，就要训练 n^2 数量级的翻译模型，即便我们将一种语言设为中间语言，所需训练的模型数也达到了 $2n$ 左右，需要耗费很大的资源；②很多低资源的语言对之间几乎没有平行语料，很难训练出高质量的神经机器翻译模型。采用中间语言的方法时，由于需要进行两次翻译，错误会在翻译路径上积累，影响最后的翻译质量。因此，研究者开始研究多语言的神经机器翻译模型，希望能用单个模型实现多种语言间的互译，并由此提高低资源语言上的翻译质量。

多语言神经机器翻译的研究主要可以分为两个阶段。在第一阶段，研究者着重于探索实现多语言翻译的模型框架；在第二阶段，模型框架基本确定，研究者开始探索改进多语言翻译性能的方法。Dong 等人[124]和 Luong 等人[125]最早开始在任务学习的框架下对多语言翻译进行研究。Dong 等人[124]尝试进行从一种源语言到多种目标语言的翻译，模型用单个编码器对源语言进行编码，对每种目标语言都使用一个独立的解码器。Luong 等人[125]实现了从多种源语言到多种目标语言的翻译，模型对每种源语言都使用一个独立的编码器，对每种目标语言都使用一个独立的解码器。Firat 等人[126]提出了基于注意力机制的多语言翻译模型。不同于文献[125]，他们令不同语言对共享注意力模块的参数，由此解决注意力模块数随语言数平方增长的问题。Lee 等人[127]提出了字符级的神经机器翻译模型，只使用单个编码器和解码器实现了多种源语言到一种目标语言的翻译。Firat 等人[128]在文献[126]的基础上，再通过构建伪平行语料进行微调，实现了零资源的机器翻译。Ha 等人[129]在源端每一个单词上拼接一个指示源端语言的符号，再在源句两端加上指示目标语言的符号，仅使用单个编码器和解码器实现了多语言互译。Johnson 等人[130]使用单个编码器和解码器，通过在源句前拼接一个特殊词来指示目标语言，在单个模型里实现了多语言的互译和零资源翻译。之后的工作基本都以文献[129-130]的方法为基础，以单个编码器和解码器对多种语言进行编码和解码。

在模型框架确定后，研究者开始探索改进多语言翻译性能的方法。Lakew 等人[131]将 Transformer 模型和 RNN 模型在多语言翻译上进行对比，证实了 Transformer 模型在多语言翻译上的能力。Blackwood 等人[132]对注意力的共享机制进行探索，提出让每种目标语言独享一套注意力参数的方法。Sachan 和 Neubig[133]也对参数共享的方式进行探索，提出了仅在解码器共享一部分参数的方法，并指出语种相似程度对多语言翻译性能的影响。Platanios 等[134]提出了一种自适应的参数共享方法，通过公有参数做线性变换来得到每种语言的参数。Lu 等人[135]不对编码器和解码器做参数共享，而是引入一个共享的中间语言模块来学习语言的公有表示。Wang 等人[136]针对一对多的翻译模型，从语言标签、位置词向量和隐状态三个角度提出了不同的参数共享策略。Wang 等人[137]对词表示进行改进，每种语言除共享的词嵌入外，也有基于 n 元组的独立词表示方法。Tan 等人[138]从知识蒸馏的角度出发，让多语言翻译模型也去学习单模型的分布。为解决不同语言语序不同的问题，Murthy 等人[139]对源端语句做重排序，使模型能更好地学到源端的公有表示。Tan 等

人[140]让模型对语言进行聚类,将所有语言分成几个大类,类内语言相似度较高,从而按类训练多语言模型。为利用多语言模型改进低资源语言上的翻译,Gu等人[141]将元学习[142]应用到多语言翻译上,希望模型能通过短时间的微调迅速提高目标语言上的翻译性能。Dabre等人[143]研究了语言相似度对低资源语种翻译性能的影响。Neubig和Hu[144]在对低资源语言进行微调时,同时引入了相似语言,从而减缓了过拟合现象。Gu等人[145]对多语言系统在零资源语言对上效果不稳定的现象进行研究,并提出预训练、反向翻译等解决方案。Wang等人[146]提出了一种数据选择算法,在给定目标句时对源端语句进行采样。随后,Wang等人[147]以训练数据和开发集的梯度点乘为奖赏,提出了自适应的数据选择算法。Arivazhagan等人[148]在大规模多语言语料上进行实验,构建了可以在102种语言间互译的多语言翻译模型。

9 模型训练

传统的神经机器翻译模型采用teacher forcing算法[149],以最大化参考译文的对数似然函数为目标进行训练。在训练时,模型的解码端接收参考译文作为输入,而在测试时,模型接收自身的输出结果作为下一步的输入。这种训练与测试间的不一致性被称为曝光误差[150],它会减弱模型对翻译过程中错误译文的鲁棒性,影响模型的翻译质量。除曝光误差之外,传统的训练方法还有其他缺陷,如训练目标不全面、翻译错误沿解码方向累积、对源端输入中的噪声很敏感、参考译文单一等。

对神经机器翻译模型训练方法的改进可以按是否用teacher forcing算法进行训练分为两类。为了消除训练中的曝光误差,模型可以不采用teacher forcing算法,而是将自身的预测结果输入到下一步。Bengio等人[151]提出schedule sampling方法,以一定概率混合参考译文词与模型预测词来作为下一步预测的输入,以减小模型在训练与预测间的差距。Zhang等人[152]对这一方法进行改进,提出word oracle和sentence oracle方法,并引入gumbel噪声以提升模型鲁棒性。实验显示,文中所提方法在RNN和Transformer模型上均有显著提升。消除曝光误差的另一思路是进行序列级训练。Ranzato等人[150]提出用序列级指标BLEU来训练机器翻译模型,并通过强化学习算法[153]克服BLEU值的离散性问题。Shen等人[154]将最小风险训练应用到神经机器翻译模型中,用BLEU、TER、NIST等指标来训练模型。Bahdanau等人[155]针对序列预测问题设计actor-critic算法,用一个critic预测整个词表上的期望奖赏值,目标为降低强化学习过程中的方差。Wu等人[9]设计单句评价指标GLEU,用强化学习方法进行序列级训练,优化生成译文的GLEU值。He等人[156]提出对偶学习方法,以反向翻译的概率为奖赏,用强化学习算法训练模型。Wu等人[157]对强化学习方法在神经机器翻译中的技巧进行了实验调研。Edunov等人[158]对序列级训练中各种损失函数的效果进行了实验验证。Shao等人[73]针对非自回归模型对强化学习方法做改进,降低了梯度估计的方差。Gu等人[159]采用deterministic policy gradient技术,实现了神经机器翻译模型直接可导的序列级训练。Shao等人[160]提出基于概率化n元组匹配的序列级训练目标,对n元组做概率化处理,使得训练目标直接可导。此外,也有人在神经机器翻译中引入gumbel softmax技术[161-163],使模型能以可导的方式从预测的概率分布中采样。

另一类改进仍采用teacher forcing算法,针对神经机器模型的其他方面的缺陷来改进训练方法。Weng等人[164]令模型直接从解码器的隐状态中预测待翻译的词袋,以此增强解码器隐状态的表示能力。Ma等人[165]在进行词级训练的同时也引入词袋损失,希望各个位置的概率分布综合起来与参考译文的词袋相近。Tu等人[166]引入重建损失,希望模型在解码时不丢失源端信息,能从解码器的隐状态中重建出源端句子。Hassan等人[12]针对单向模型错误累积的问题,提出对左到右(L2R)和右到左(R2L)的翻译模型的KL一致性约束,令双向模型能够互相学习。Yang等人[167]提出句级一致性损失,希望源端句子在编码后的表示能与目标端句子的词嵌入尽量接近。针对神经机器翻译模型对输入噪声敏感的问题,一系列工作在训练时就在源端添加噪声,以提升模型的鲁棒性[168-170]。Chousa等人[171]对译文单词的0-1分布做平滑,不仅最大化参考译文对应词的似然概率,也根据词嵌入信息对相近词给予一定权重。Feng等人[172]针对参考译文单词的0-1分布,引入一个用于评估的解码器,令翻译解码器同时也以最小化与评估解码器的差异为目标进行训练。Norouzi等人[173]针对参考译文单一的问题,求出一个以参考译文为中心的概率分布,并从该概率分布中采样,最大化模型在采样得到的译文

上的对数似然概率。Elbayad 等人[174]对上述方法做扩展，借助重要性采样技术，使之适用于任意的距离函数(如 BLEU 值)。另外，该工作还以词嵌入相似性为距离指标，将该方法扩展到了词级上。

10　结语

神经机器翻译目前已成为机器翻译的主流方法，受到研究者的广泛关注。本文首先对神经机器翻译模型的发展流程和其存在的问题做了简要介绍，随后回顾了神经机器翻译主流框架的演变，并对同声传译、非自回归模型、多模态机器翻译、篇章翻译、领域自适应、多语言翻译、模型训练等方向的前沿进展做了简要介绍。

参考文献

[1] Brown P F, Cocke J, Della Pietra S A, et al. A statistical approach to machine translation[J]. Computational Linguistics, 1990, 16(2): 79-85.

[2] Brown P F, Pietra V J D, Pietra S A D, et al. The mathematics of statistical machine translation: Parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.

[3] Kalchbrenner N, Blunsom P. Recurrent continuous translation models[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1700-1709.

[4] Sutskever I, Vinyals O, Le Q V.Sequence to sequence learning with neural networks[C]//Procedings of Advances in Neural Information Processing Systems, 2014: 3104-3112.

[5] Cho K, Gulcehre B M C, Bahdanau D, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 1406.1078, 2014.

[6] Cho K, van Merrienboer B, Bahdanau D, et al. On the properties of neural machine translation: encoder-decoder approaches[C]//Proceedings of SSST-8, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014: 103-111.

[7] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409.0473, 2014.

[8] Junczys-Dowmunt M, Dwojak T, Hoang H.Is neural machine translation ready for deployment? A case study on 30 translation directions[J]. arXiv preprint arXiv: 1610.01108, 2016.

[9] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv preprint arXiv: 1609.08144, 2016.

[10] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: 1243-1252.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Procedings of Advances in Neural Information Processing Systems, 2017: 5998-6008.

[12] Hassan H, Aue A, Chen C, et al. Achieving human parity on automatic chinese to english news translation[J]. arXiv preprint arXiv: 1803.05567, 2018.

[13] Wu F, Fan A, Baevski A, et al. Pay less attention with lightweight and dynamic convolutions[J]. arXiv preprint arXiv: 1901.10430, 2019.

[14] Dehghani M, Gouws S, Vinyals O, et al. Universal transformers[J]. arXiv preprint arXiv: 1807.03819, 2018.

[15] So D, Le Q, Liang C. The evolved transformer[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 5877-5886.

[16] Meng F, Zhang J. DTMT: A novel deep transition architecture for neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 224-231.

[17] Wang Q, Li B, Xiao T, et al. Learning deep transformer models for machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1810-1822.

[18] Chen K, Wang R, Utiyama M, et al. Neural machine translation with reordering embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1787-1799.

[19] Dou Z Y, Tu Z, Wang X, et al.Dynamic layer aggregation for neural machine translation with routing-by-agreement[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 86-93.

[20] Sennrich R, Haddow B, Birch A. Improving meural machine translation models with monolingual data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 86-96.

[21] Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 489-500.

[22] Junczys-Dowmunt M. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data[C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 425-430.

[23] Junczys Dowmunt M. Dual conditional cross-entropy filtering of noisy parallel corpora[C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 888-895.

[24] Song K, Tan X, Qin T, et al. MASS: masked sequence to sequence pre-training for language generation[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 5926-5936.

[25] Meng F, Lu Z, Wang M, et al. Encoding source language with convolutional neural network for machine translation [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 20-30.

[26] Gehring J, Auli M, Grangier D, et al. A convolutional encoder model for neural machine translation[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 123-135.

[27] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks [C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: 933-941.

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[29] Ba J L, Kiros J R, Hinton G E. Layer normalization [J]. arXiv preprint arXiv: 1607.06450, 2016.

[30] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting on association for Computational Linguistics. Association for Computational Linguistics, 2002: 311-318.

[31] Cho K, Esipova M. Can neural machine translation do simultaneous translation? [J]. arXiv preprint arXiv: 1606.02012, 2016.

[32] Gu J, Neubig G, Cho K, et al. Learning to translate in real-time with neural machine translation [C]// Proceedings of 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Association for Computational Linguistics (ACL), 2017: 1053-1062.

[33] Ma M, Huang L, Xiong H, et al. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3025--3036.

[34] Arivazhagan N, Cherry C, Macherey W, et al. Monotonic infinite lookback attention for simultaneous machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1313-1323.

[35] Dalvi F, Durrani N, Sajjad H, et al. Incremental decoding and training methods for simultaneous translation in neural machine translation [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 493-499.

[36] Zheng B, Zheng R, Ma M, et al. Simultaneous translation with flexible policy via restricted imitation learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5816-5822.

[37] Zheng B, Zheng R, Ma M, et al. Simpler and faster learning of adaptive policies for simultaneous 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), 2019: 1349-1354.

[38] Ma X, Pino J, Cross J, et al. Monotonic multihead attention [J]. arXiv preprint arXiv: 1909.12406, 2019.

[39] Alinejad A, Siahbani M, Sarkar A. Prediction improves simultaneous neural machine translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3022-3027.

[40] Zheng R, MaM, Zheng B, et al. Speculative beam search for simultaneous 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), 2019: 1395-1402.

[41] Specia L, Frank S, Sima'an K, et al. A shared task on multimodal machine translation and crosslingual image description[C]//Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers, 2016: 543-553.

[42] Elliott D, Frank S, Barrault L, et al. Findings of the second shared task on multimodal machine mranslation and multilingual image description[C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 215-233.

[43] Barrault L, Bougares F, Specia L, et al. Findings of the third shared task on multimodal machine translation[C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 304-323.

[44] Caglayan O, Aransa W, Bardet A, et al. LIUM-CVC submissions for WMT17 multimodal translation task [C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 432-439.

[45] Libovický J, Helcl J, Tlustý M, et al. CUNI system for WMT16 automatic post-editing and multimodal translation tasks[C]//Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers, 2016: 646-654.

[46] Calixto I, Liu Q, Campbell N. Incorporating global visual features into attention-based neural machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 992-1003.

[47] Huang P Y, Liu F, Shiang S R, et al. Attention-based multimodal neural machine translation[C]//Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers, 2016: 639-645.

[48] Elliott D, Frank S, Hasler E. Multilingual image description with neural sequence models[J]. arXiv preprint arXiv: 1510.04709, 2015.

[49] Madhyastha P S, Wang J, Specia L. Sheffield multiMT: using object posterior predictions for multimodal machine translation[C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 470-476.

[50] Caglayan O, Barrault L, Bougares F. Multimodal attention for neural machine translation[J]. arXiv preprint arXiv: 1609.03976, 2016.

[51] Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1913-1924.

[52] Delbrouck J B, Dupont S. An empirical study onthe effectiveness of images in multimodal neural machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 910-919.

[53] Libovický J, Helcl J. Attention strategies for multi-source sequence-to-sequence learning[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 196-202.

[54] Zhou M, Cheng R, Lee Y J, et al. A visual attention grounding neural model for multimodal machine translation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3643-3653.

[55] Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness[J]. arXiv preprint arXiv: 1906.07701, 2019.

[56] Toyama J, Misono M, Suzuki M, et al. Neural machine translation with latent semantic of image and text[J]. arXiv preprint arXiv: 1611.08459, 2016.

[57] Calixto I, Rios M, Aziz W. Latent variable model for multi-modal translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6392-6405.

[58] Pengcheng Y, Boxing C, Pei Z, et al. Visual agreement regularized training for multi-modal machine translation[J]. arXiv preprint arXiv: 1912.12014, 2019.

[59] Gu J, Bradbury J, Xiong C, et al. Non-autoregressive neural machine translation[J]. arXiv preprint arXiv: 1711.02281, 2017.

[60] Shao C, Feng Y, Zhang J, et al. Retrieving sequential information for non-autoregressive neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3013-3024.

[61] Kim Y, Rush A M. Sequence-level knowledge distillation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 1317-1327.

[62] Zhou C, Neubig G, Gu J. Understanding knowledge distillation in non-autoregressive machine translation[J]. arXiv preprint arXiv: 1911.02727, 2019.

[63] Wei B, Wang M, Zhou H, et al. Imitation learning for non-autoregressive neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1304-1312.

[64] Guo J, Tan X, Xu L, et al. Fine-tuning by curriculum learning for non-autoregressive neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[65] Li Z, Lin Z, He D, et al. Hint-based training for non-autoregressive 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), 2019: 5708-5713.

[66] Kaiser L, Bengio S, Roy A, et al. Fast decoding in sequence models using discrete latent variables[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 2395-2404.

[67] Roy A, Vaswani A, Neelakantan A, et al. Theory and experiments on vector quantized autoencoders[J]. arXiv preprint arXiv: 1805.11063, 2018.

[68] Ma X, Zhou C, Li X, et al. FlowSeq: Non-autoregressive conditional sequence generation with generative flow[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), 2019: 4273-4283.

[69] Shu R, Lee J, Nakayama H, et al. Latent-variable non-autoregressive neural machine translation with

deterministic inference using a delta posterior[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[70] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv: 1312.6114, 2013.

[71] Bao Y, Zhou H, Feng J, et al. Non-autoregressive transformer by position learning[J]. arXiv preprint arXiv: 1911.10677, 2019

[72] Ran Q, Lin Y, Li P, et al. Guiding non-autoregressive neural machine translation decoding with reordering information[J]. arXiv preprint arXiv: 1911.02215, 2019.

[73] Wang Y, Tian F, He D, et al. Non-autoregressive machine translation with auxiliary regularization [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[74] Sun Z, Li Z, Wang H, et al. Fast structured decoding for sequence models[C]//Procedings of Advances in Neural Information Processing Systems, 2019: 3011-3020.

[75] Libovický J, Helcl J. End-to-end non-autoregressive neural machine translation with connectionist temporal classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3016-3021.

[76] Shao C, Zhang J, Feng Y, et al. Minimizing the bag-of-ngrams difference for non-autoregressive neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[77] Lee J, Mansimov E, Cho K. Deterministic non-autoregressive neural sequence modeling by iterative refinement[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 1173-1182.

[78] Ghazvininejad M, LevyO, Liu Y, et al. Mask-predict: parallel decoding of conditional masked language models[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), 2019: 6114-6123.

[79] Gu J, Wang C, Zhao J. Levenshtein transformer [C]//Procedings of the Neural Information Processing Systems, 2019: 11181--11191.

[80] VoitaE, Sennrich R, Titov I. When a good translation is wrong in context: context-aware machine translation improves on deixis, ellipsis, and lexical cohesion[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1198-1212.

[81] Jean S, Lauly S, Firat O, et al. Does neural machine translation benefit from larger context? [J]. arXiv preprint arXiv: 1704.05135, 2017.

[82] Bawden R, Sennrich R, Birch A, et al. Evaluating discourse phenomena in neural machine translation [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 1304-1313.

[83] Wang L, Tu Z, Way A, et al. Exploiting cross-sentence context for neural machine translation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2826-2831.

[84] Voita E, Serdyukov P, Sennrich R, et al. Context-aware neural machine translation learns anaphora resolution[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1264-1274.

[85] Zhang J, Luan H, Sun M, et al. Improving the transformer translation model with document-level context[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 533-542.

[86] Miculicich L, Ram D, Pappas N, et al. Document-level neural machine translation with hierarchical attention networks[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2947-2954.

[87] Maruf S, Martins A F T, Haffari G. Selective attention for context-aware neural machine translation [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3092-3102.

[88] Yang Z, Zhang J, Meng F, et al. Enhancing context modeling with a query-guided capsule network for document-level 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), 2019: 1527-1537.

[89] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C]//Procedings of Advances in Neural Information Processing Systems, 2017: 3856-3866.

[90] Sabour S, Frosst N, Hinton G. Matrix capsules with EM routing[C]//Proceedings of the 6th International Conference on Learning Representations. ICLR, 2018: 1-15.

[91] Tu Z, Liu Y, Shi S, et al. Learning to remember translation history with a continuous cache[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 407-420.

[92] Kuang S, Xiong D, Luo W, et al. Modeling coherence for neural machine translation with dynamic and topic caches[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 596-606.

[93] Maruf S, Haffari G. Document context neural machine translation with memory networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1275-1284.

[94] Xiong H, He Z, Wu H, et al. Modeling coherence for discourse neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 7338-7345.

[95] Bosselut A, Celikyilmaz A, He X, et al. Discourse-aware neural rewards for coherent text generation [C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 173-184.

[96] Voita E, Sennrich R, Titov I. Context-aware monolingual repair 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), 2019: 876-885.

[97] Moore R C, Lewis W. Intelligent selection of language model training data[C]//Proceedings of the ACL 2010 Conference. Association for Computational Linguistics, 2010: 220-224.

[98] Axelrod A, He X, Gao J. Domain adaptation via pseudo in-domain data selection[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 355-362.

[99] Wang R, Finch A, Utiyama M, et al. Sentence embedding for neural machine translation domain adaptation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 560-566.

[100] van der Wees M, Bisazza A, Monz C. Dynamic data selection for neural machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1400-1410.

[101] Hu J, Xia M, Neubig G, et al. Domain adaptation of neural machine translation by lexicon induction [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2989-3001.

[102] Chu C, Dabre R. Multilingual multi-domain adaptation approaches for neural machine translation[J]. arXiv preprint arXiv: 1906.07978, 2019.

[103] Imankulova A, DabreR, Fujita A, et al. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation[C]//Proceedings of Machine Translation Summit XVII Volume 1: Research Track, 2019: 128-139.

[104] Freitag M, Al-Onaizan Y. Fast domain adaptation for neural machine translation[J]. arXiv preprint arXiv: 1612.06897, 2016.

[105] Luong M T, Manning C D. Stanford neural machine translation systemsfor spoken language domains [C]//Proceedings of the International Workshop on Spoken Language Translation, 2015: 76-79.

[106] Servan C, Crego J, Senellart J. Domain specialization: A post-training domain adaptation for neural machine translation[J]. arXiv preprint arXiv: 1612.06141, 2016.

[107] Dakwale P, Monz C. Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data[C]//Proceedings of the XVI Machine Translation Summit, 2017: 117.

[108] Chu C, Dabre R, Kurohashi S. An empirical comparison of domain adaptation methods for neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 385-391.

[109] Barone A V M, Haddow B, Germann U, et al. Regularization techniques for fine-tuning in neural machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1489-1494.

[110] Wang R, Utiyama M, Liu L, et al. Instance weighting for neural machine translation domain adaptation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1482-1488.

[111] Chen B, Cherry C, Foster G, et al. Cost weighting for neural machine translation domain adaptation [C]//Proceedings of the 1st Workshop on Neural Machine Translation, 2017: 40-46.

[112] Yan S, Dahlmann L, Petrushkov P, et al. Word-based domain adaptation for neural machine translation[J]. arXiv preprint arXiv: 1906.03129, 2019.

[113] Vilar D. Learning hidden unit contribution for adapting neural machine translation models[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2, 2018: 500-505.

[114] Zhang X, Shapiro P, Kumar G, et al. Curriculum learning for domain adaptation in neural machine translation[C]//Proceedings of the 2019 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,, 2019: 1903-1915.

[115] Zeng J, Liu Y, Lu Y, et al. Iterative dual domain 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), 2019: 845-854.

[116] Gulcehre C, Firat O, Xu K, et al. On using monolingual corpora in neural machine translation[J]. arXiv preprint arXiv: 1503.03535, 2015.

[117] Dou ZY, Wang X, Hu J, et al. Domain differential 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), 2019: 59-69.

[118] Khayrallah H, Kumar G, Duh K, et al. Neural lattice search for domain adaptation in machine translation[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017: 20-25.

[119] Britz D, Le Q, Pryzant R. Effective domain mixing for neural machine translation[C]//Proceedings of the 2nd Conference on Machine Translation, 2017: 118-126.

[120] Kobus C, Crego J, Senellart J. Domain control for neural machine translation[C]//Proceedings of the International Conference Recent Advances in Natural Language Processing. RANLP 2017, 2017: 372-378.

[121] Thompson B, Khayrallah H, Anastasopoulos A, et al. Freezing subnetworks to analyze domain adaptation in neural machine translation[C]//Proceedings of the 3rd Conference on Machine Translation: Research Papers, 2018: 124-132.

[122] Wuebker J, Simianer P, DeNero J. Compact personalized models for neural machine translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 881-886.

[123] Gu S, Feng Y, Liu Q. Improving domain adaptation translation with domain invariant and specific information[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3081-3091.

[124] Dong D, Wu H, He W, et al. Multi-task learning for multiple language translation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 1723-1732.

[125] Luong M T, Le Q V, Sutskever I, et al. Multi-task sequence to sequence learning[J]. arXiv preprint arXiv: 1511.06114, 2015.

[126] Firat O, Cho K, Bengio Y. Multi-Way, Multilingual neural machine translation with a shared attention mechanism[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 866-875.

[127] Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 365-378.

[128] Firat O, Sankaran B, Al-Onaizan Y, etal. Zero-resource translation with multi-lingual neural machine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 268-277.

[129] Ha T L, Niehues J, Waibel A. Toward multilingual neural machine translation with universal encoder and decoder[J]. arXiv preprint arXiv: 1611.04798, 2016.

[130] 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.

[131] Lakew S M, Cettolo M, Federico M. A comparison of transformer and recurrent neural networks on multilingual neural machine translation[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 641-652.

[132] Blackwood G, Ballesteros M, Ward T. Multilingual neural machine translation with task-specific attention[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 3112-3122.

[133] Sachan D, Neubig G. Parameter sharing methods for multilingual self-attentional translation Models [C]//Proceedings of the 3rd Conference on Machine Translation: Research Papers, 2018: 261-271.

[134] Platanios E A, SachanM, 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, 2018: 425-435.

[135] Lu Y, Keung P, Ladhak F, et al. A neural interlingua for multilingual machine translation[C]//Proceedings of the 3rd Conference on Machine Translation: Research Papers, 2018: 84-92.

[136] Wang Y, Zhang J, Zhai 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, 2018: 2955-2960.

[137] Wang X, Pham H, Arthur P, et al. Multilingual neural machine translation with soft decoupled encoding[J]. arXiv preprint arXiv: 1902.03499, 2019.

[138] Tan X, Ren Y, He D, et al. Multilingual neural machine translation with knowledge distillation[J]. arXiv preprint arXiv: 1902.10461, 2019.

[139] Murthy R, Kunchukuttan A, Bhattacharyya P. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3868-3873.

[140] Tan X, Chen J, 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), 2019: 962-972.

[141] GuJ, Wang Y, Chen Y, et al. Meta-learning for low-resource neural machine translation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3622-3631.

[142] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fastadaptation of deep networks[C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1126-1135.

[143] Dabre R, Nakagawa T, Kazawa H. An empirical study of language relatedness for transfer learning in neural machine translation[C]//Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, 2017: 282-286.

[144] Neubig G, Hu J. Rapid adaptation of neural machine translation to new languages[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 875-880.

[145] Gu J, 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, 2019: 1258-1268.

[146] Wang X, Neubig G. Target conditioned sampling: optimizing data selection for multilingual neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5823-5828.

[147] Wang X, PhamH, Michel P, et al. Optimizing data usage via differentiable rewards[J]. arXiv preprint arXiv: 1911.10088, 2019.

[148] Arivazhagan N, Bapna A, Firat O, et al. Massively multilingual neural machine translation in the wild: Findings and challenges[J]. arXiv preprint arXiv: 1907.05019, 2019.

[149] Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks[J]. Neural Computation, 1989, 1(2): 270-280.

[150] Ranzato M A, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks[J]. arXiv preprint arXiv: 1511.06732, 2015.

[151] Bengio S, Vinyals O, Jaitly N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[C]//Procedings of Advances in Neural Information Processing Systems, 2015: 1171-1179.

[152] Zhang W, FengY, Meng F, et al. Bridging the gap between training and inference for neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4334-4343.

[153] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3-4): 229-256.

[154] Shen S, Cheng Y, He Z, et al. Minimum risk training for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1683-1692.

[155] Bahdanau D, Brakel P, Xu K, et al. An actor-critic algorithm for sequence prediction[J]. arXiv preprint arXiv: 1607.07086, 2016.

[156] He D, Xia Y, Qin T, et al. Dual learning for machine translation[C]//Procedings of Advances in Neural Information Processing Systems, 2016: 820-828.

[157] Wu L, Tian F, Qin T, et al. A study of reinforcement learning for neural machine translation[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3612-3621.

[158] Edunov S, Ott M, Auli M, et al. Classical structured prediction losses for sequence to sequence learning[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 355-364.

[159] Gu J, Cho K, Li V O K. Trainable greedy decoding for neural machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1968-1978.

[160] Shao C, Chen X, Feng Y. Greedy search with probabilistic n-gram matching for neural machine translation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4778-4784.

[161] Gu J, Im D J, Li V O K. Neural machine translationwith gumbel-greedy decoding[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[162] Niu X, Xu W, Carpuat M. Bi-directional differentiable input reconstruction for low-resource neural machine translation[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 442-448.

[163] Xu W, Niu X, Carpuat M. Differentiable sampling with flexible reference word order for neural machine translation[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 2047-2053.

[164] Weng R, Huang S, Zheng Z, et al. Neural machine translation with word predictions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 136-145.

[165] Ma S, SUN Xu, Wang Y, et al. Bag-of-words as target for neural machine translation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 332-338.

[166] Tu Z, Liu Y, Shang L, et al. Neural machine translation with reconstruction[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[167] Yang M, Wang R, Chen K, et al. Sentence-level agreement for neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3076-3082.

[168] Sperber M, Niehues J, Waibel A. Toward robust neural machine translation for noisy input sequences [C]//Pcoceedigns of the International Workshop on Spoken Language Translation (IWSLT), 2017.

[169] Sano M, Suzuki J, Kiyono S. Effective adversarial regularization for neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 204-210.

[170] ChengY, Jiang L, Macherey W. Robust neural machine translation with doubly adversarial inputs [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4324-4333.

[171] Chousa K, Sudoh K, Nakamura S. Training neural machine translation using word embedding-based loss[J]. arXiv preprint arXiv: 1807.11219, 2018.

[172] Feng Y, Xie W, Gu S, et al. Modeling fluency and faithfulness for diverse neural machine translation [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[173] Norouzi M, Bengio S, Jaitly N, et al. Reward augmented maximum likelihood for neural structured prediction[C]//Procedings of Advances in Neural Information Processing Systems, 2016: 1723-1731.

[174] Elbayad M, Besacier L, Verbeek J. Token-level and sequence-level loss smoothing for RNN language models[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2094-2103.

冯洋(1982—),博士,副研究员,博士生导师,主要研究领域为机器翻译。
E-mail: yangyang@ict.ac.cn

邵晨泽(1996—),博士研究生,主要研究领域为机器翻译。
E-mail: shaochenze18z@ict.ac.cn