马皓明, 陈永富, 李付学,等. 基于助教的知识蒸馏增强神经机器翻译方法[J]. 智能计算机与应用,2026,16(4):219-225.
DOI:10.20169/j.issn.2095-2163.24061104

基于助教的知识蒸馏增强神经机器翻译方法

马皓明[1,2], 陈永富[1,2], 李付学[3], 闫 红[3]
(1 沈阳化工大学 计算机科学与技术学院, 沈阳 110142; 2 辽宁省化工过程工业智能化技术重点实验室, 沈阳 110142
3 营口理工学院 电气工程学院, 辽宁 营口 115014)

摘 要: 神经机器翻译(NMT)大模型在许多翻译任务中取得了卓越的成绩。同时研究人员引入了知识蒸馏方法,通过教师(大)模型指导学生(小)模型训练,以提升学生模型的性能,从而实现大模型向小模型的性能迁移。然而实验结果表明,当学生模型和教师模型之间的规模差距过大时,学生模型的性能会下降。为解决这一问题,本文提出了一种多步蒸馏策略,即在知识蒸馏过程中引入规模介于教师模型和学生模型之间的助教(Teaching Assistant,TA)模型,该模型由教师模型通过知识蒸馏训练得来,同时用来指导学生模型的训练,以进一步提升学生模型的性能。本文在 WMT14 和 WMT16 机器翻译任务上的实验结果表明,与基线模型和现有知识蒸馏方法相比,基于助教的知识蒸馏方法(ATKD)能够有效地提高模型的翻译性能。

关键词: 助教模型; 神经机器翻译; 知识蒸馏; 多步蒸馏; 性能转移
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0219-07

Knowledge distillation-enhanced neural machine translation method based on teaching assistant

MA Haoming[1,2], CHEN Yongfu[1,2], LI Fuxue[3], YAN Hong[3]
(1 College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China;
2 Liaoning Key Laboratory of Intelligent Technology for Chemical Process Industry, Shenyang 110142, China;
3 College of Electrical Engineering, Yingkou Institute of Technology, Yingkou 115014, Liaoning, China)

Abstract: Large models of neural machine translation (NMT) have achieved excellent results in various translation tasks. Then researchers introduce a knowledge distillation method, through the teacher (large) model to guide the student (small) model training, so that the performance of the student model is improved, so as to achieve performance transfer from large to small models. However, the experimental results show that when the scale gap between the student model and the teacher model is too large, the performance of the student model will decline. In order to solve this problem, a multi-step distillation strategy is proposed in this paper, that is, a teaching assistant model whose scale is between the teacher model and the student model is introduced into the knowledge distillation process. The teaching assistant model is obtained from the teacher model through the knowledge distillation training, and is used to guide the training of the student model, so as to further improve the performance of the student model. The experimental results of WMT14 and WMT16 machine translation tasks show that compared with the baseline model and existing knowledge distillation methods, the teaching assistant-based knowledge distillation method (ATKD) can effectively improve the translation performance of the model.

Key words: Teaching Assistant Model; Neural Machine Translation; knowledge distillation; multistep distillation; performance transfer

0 引 言

近年来,神经机器翻译(Neural Machine Translaton, NMT)已逐渐成为机器翻译领域的主流范式,其中最具代表性的模型是 2017 年由 Google 提出的 Transformer 模型[1]。该模型通过自注意力机制和残

基金项目: 辽宁省自然科学基金区域联合基金(2022-YKLH-18);营口理工学院校级科研项目(ZDIL202305,YBL202316)。
作者简介: 马皓明(2000—),男,硕士研究生,主要研究方向:机器翻译;陈永富(1999—),男,硕士研究生,主要研究方向:自然语言处理;闫红(1984—),女,硕士,副教授,主要研究方向:自然语言处理。
通信作者: 李付学(1985—),男,博士,副教授,主要研究方向:自然语言处理。Email:lifuxue@yku.edu.cn。
收稿日期: 2024-06-11

差连接等技术,实现对输入序列的深度理解和表征,在众多机器翻译任务中取得了卓越的性能。然而随着研究的推进,基于 Transformer 模型的神经机器翻译模型在性能上不断提升的同时,其规模也日益庞大。这些大模型的训练和部署消耗大量计算资源,且难以直接应用于流行的嵌入式和移动设备。因此,如何高效、便捷地实现大模型的性能迁移与轻量化部署成为神经机器翻译领域亟待解决的关键问题。

知识蒸馏[2]作为一种新兴的通用模型压缩与迁移学习方法,近年来展现出蓬勃的发展活力。其核心目标是将复杂的深度学习模型(教师模型)的知识迁移到简单的模型(学生模型),使学生模型能够模拟教师模型的输出结果,从而实现模型规模压缩与小模型性能提升的效果。文献[3]中提出的层间知识蒸馏方法旨在将高层网络的抽象知识迁移到低层网络,以提升整个模型的翻译质量。与传统的教师模型和学生模型之间的知识蒸馏不同,层间知识蒸馏实现的是同一模型内不同层之间的知识传递。此外,文献[4]还探讨了利用知识蒸馏方法压缩多语言神经机器翻译(Multilingual Neural Machine Translation, MNMT)模型。文献[5]提出了基于递进式半知识蒸馏的方法。

然而,随着神经机器翻译模型规模的不断扩大,当教师模型和学生模型的规模相差过大时,教师模型的输出 logit(对输入信息的预测)中包含的 soft 信息(包括负标签信息)会减少,导致学生模型可学到的内容减少,进而影响知识蒸馏的性能。目前,神经机器翻译领域对于知识蒸馏的研究大多针对教师模型和学生模型之间传递的具体知识,例如从“知识”入手,通过分析传递知识的具体内容,重新分配权重,以此来改进知识蒸馏的效果等。

针对神经机器翻译领域内对知识蒸馏模型层面研究的不足,本文从参与知识蒸馏的模型角度入手,提出了一种新的神经机器翻译知识蒸馏框架。该框架引入助教模型,以解决模型规模差距过大导致性能下降的问题。助教模型由教师模型训练得到,其规模介于教师模型和学生模型之间。随后,助教模型担任教师的角色对学生模型进行训练。因此,相较于直接拟合教师模型,学生模型能更好地拟合助教模型的输出。

实验结果表明,引入助教模型后的神经机器翻译模型在 WMT14 En-De、WMT14 En-Fr、WMT16 En-Ro 及 WMT 16 Tr-En 数据集上的知识蒸馏性能优于基线模型。因此,该方法的引入对神经机器翻译模型的压缩和性能提升具有积极作用。

1 知识蒸馏方法

深度学习在学术界和工业界都获得了巨大发展,核心原因在于其优异的可扩展性与处理大量信息的高效编码技术。然而,受限于计算资源是深度学习面临的核心挑战之一,大型深度学习模型难以直接部署在资源受限的设备(如嵌入式设备和移动设备)上。因此,大量的模型压缩和加速技术应运而生,知识蒸馏便是其中的代表。

在神经机器翻译领域,知识蒸馏具有以下 4 个作用:首先,用于对大规模翻译模型进行压缩。如,基于 Transformer 的高性能机器翻译模型在资源受限的环境中部署较为困难,而知识蒸馏可以有效地解决这一问题,构建出更小、更轻量级的模型,同时保持较高的性能水平。其次,知识蒸馏有助于提高模型性能。通过知识蒸馏,学生模型可以从教师模型中学习到更好的泛化能力,从而提升性能。因此,研究人员可以充分利用开源大模型来训练自己的模型,提高研究效率。第三,知识蒸馏还能降低模型的推理时延。在某些应用场景中,模型的推理速度至关重要。通过将高时延教师模型的知识迁移到低时延学生模型中,可以降低整体推理时延。最后,知识蒸馏对于跨数据集域的知识迁移非常有用。其可以将不同数据集中的知识进行集成和迁移,从而在不同领域或任务之间共享知识。

Wang[6]等设计了一种新颖的协议,可以有效地分析不同样本之间的影响。发现教师模型的知识并不是越多越好,特定样本上的知识甚至可能损害整体蒸馏性能。为了解决这些问题,提出了两种简单而有效的策略:批次级别和全局级别的样本选择,以选择适合蒸馏的样本。Liu[7]等首次将自蒸馏和自适应推理引入到 NLP 模型中,并提出了一个快速版本的 BERT,即 FastBERT。具体来说,FastBERT 在训练阶段采用自蒸馏机制,在推理阶段采用自适应机制,以更小的精度损失获得更高的效率。Jiang 等[8]提出将多个单语言模型(教师模型)的结构知识提取到统一的多语言模型(学生)中,以缩小单语模型与统一多语言模型之间的差距,同时提出了在序列标记中将单语言模型的知识提取到单个多语言模型的两种结构级方法:Top-K 知识蒸馏和后验蒸馏。Zhang[9]蒂在图神经网络的知识蒸馏实践中,发现现有的基于单一 GNN 模型的单教师 GNN 知识蒸馏方法并不理想,为此提出了一种提炼多尺度知识

的新方法，即从具有不同层数的多个 GNN 教师模型中学习。上述研究为本文方法的提出提供了理论依据。Mirzadeh[10] 等在卷积神经网络中的研究证明，随着教师模型卷积神经网络层数的增加，经过知识蒸馏后的学生模型表现呈现先上升后下降的趋势。这表明教师模型的规模并非越大越好，即教师模型与学生模型性能差距达到一定程度后，知识蒸馏的效果会降低。Jafari[11] 等的实验结果证明，知识蒸馏过程中传递的信息源自于教师模型的知识。然而，当教师模型性能足够强大时，这部分知识的丰富程度下降，即教师模型预测的输出中正确的比率很大，导致预测结果中负标签的信息很少，从而影响知识蒸馏的效率。

2　神经机器翻译

2.1　神经机器翻译方法

神经机器翻译（NMT）是一种利用深度神经网络实现自动翻译[12] 的方法，同时注意力机制的引入解决了长距离依赖问题，进一步提高了翻译性能。该方法通过编码器-解码器结构将源语言句子转换为目标语言句子。具体而言，编码器将输入的源语言句子编码成一个固定长度的向量，而解码器则将该向量解码成目标语言句子。神经机器翻译已经在实际应用中取得了显著的成效。

2.2　Transformer 模型介绍

Transformer 是一种利用注意力机制[13] 来提高训练速度和性能的模型。其具有并行化计算的优势，因此在模型的复杂程度和性能上都表现出更高的水平。标准的 Transformer 模型结构如图 1 所示，由编码器和解码器组成，编码器和解码器均由一个编码层和若干相同的 Transformer 模块层构成。当以基于 Transformer 的神经机器翻译模型为基础时，其性能通常较高，但模型规模也随之增大，导致部署和调用的难度增加。

Transformer 模型的自注意力机制（Self-Attention）和前馈神经网络赋予其强大的特征表示能力。自注意力机制有效解决了机器翻译等任务中存在的长距离依赖问题，因此在处理语言文字序列类型任务时非常适用。自注意力机制是 Transformer 模型的关键组成部分，与传统的注意力机制[14] 相比，自注意力机制对内部信息之间的关联性进行了进一步挖掘，因此 Transformer 模型能够更精准地捕捉词与词之间的特征。自注意力机制中，首先会对如下 3 个矩阵进行运算：查询（Query，$\boldsymbol{Q}$）、键（Key，$\boldsymbol{K}$）和值（Value，$\boldsymbol{V}$）。然后进行缩放点积操作，公式如下：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{1}$$

多头注意力机制则采用 h 个注意力头表示输入信息，将多头注意力的输出拼接乘以权重矩阵得到向量输出，公式如下：

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)\,W^O \tag{2}$$

$$\mathrm{head}_i = \mathrm{Attention}(\boldsymbol{Q}W_i^{\boldsymbol{Q}},\boldsymbol{K}W_i^{\boldsymbol{K}},\boldsymbol{V}W_i^{\boldsymbol{V}}) \tag{3}$$

其中，$W_i^{\boldsymbol{Q}} \in R^{d_{\mathrm{model}}\times d_k}$，$W_i^{\boldsymbol{K}} \in R^{d_{\mathrm{model}}\times d_k}$，$W_i^{\boldsymbol{V}} \in R^{d_{\mathrm{model}}\times d_k}$，$W_i^O \in R^{hd_v\times d_k}$。

为了更好地对深层网络结构进行优化，Transformer 在每个子层之中使用了残差连接和层正则化。公式如下：

$$\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x)) \tag{4}$$

与编码器相比，解码器同样堆叠了多个相同的结构。除了编码器中的两个子层，解码器还额外包含一个编码器-解码器注意力子层。该子层对编码器的输出向量再次执行注意力加权操作，从而捕获当前翻译与特征向量之间的联系。

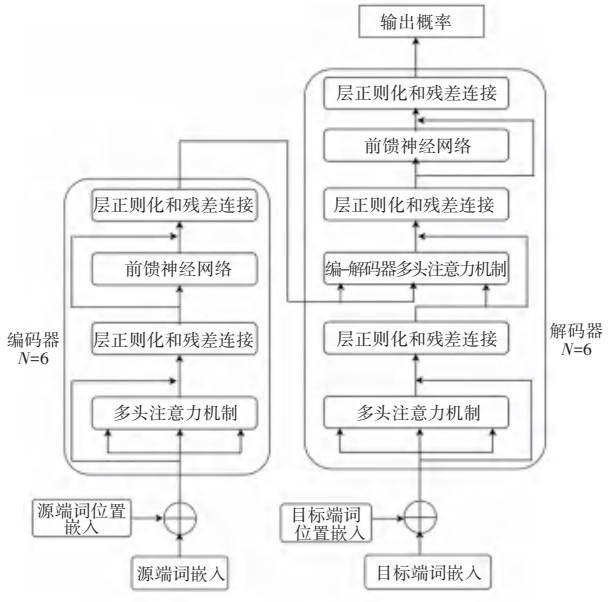

图 1　Transformer 模型结构图

Fig. 1　Transformer model structure diagram

3　助教增强知识蒸馏方法

本文引入助教模型来解决教师模型规模过大导致的知识蒸馏性能下降问题，从而提升知识蒸馏对神经机器翻译的增强效果。其中助教的规模和能力介于教师和学生之间。具体而言，助教模型由教师

模型通过知识蒸馏训练得到,然后扮演教师的角色,通过知识蒸馏来训练学生模型。

该方法包含 3 个步骤。首先,使用 fairseq[15] 工具包对 Transformer_vaswani_wmt_big 模型进行训练,将其作为教师模型。为了对其进行充分训练,实验中设置训练步数 num_updates 为 300 000。然后,在相同数据集上由教师模型通过知识蒸馏训练得到助教模型,其中,助教模型基于 Transformer 模型,通过设置输入词向量维度、前馈神经网络维度等参数来控制其规模介于教师模型和学生模型之间,从而实现助教模型对教师模型和学生模型之间存在的"空白"的填充。值得注意的是,本文中的学生模型默认为 Transformer_base。Transformer_base 的结构为六层编码器、六层解码器,输入词向量维度为 512,前馈神经网络维度为 2 048。最后,由助教模型扮演"教师"角色,对学生模型进行知识蒸馏。为确保实验结果的可靠性,保证第二步和第三步中知识蒸馏的损失函数和相应参数值的一致性。然后,对助教模型训练得到的学生模型进行性能测试。为验证本文方法的有效性,测试所用的数据集均为领域内的公开大规模数据集。

知识蒸馏的核心思想在于让学生模型不仅通过真实标签提供的信息进行训练,还要通过观察教师模型输出的方式进行训练、表示和处理数据。其中,a_t 和 a_s 分别表示教师模型和学生模型的 logit(最终 Softmax 的输入)。在经典的监督学习中,学生模型 Softmax(a_s) 的输出与真实值标签 y_r 之间的不匹配通常使用交叉熵损失来进行惩罚,公式如下:

$$L_{SL} = H(\text{Softmax}(a_s), y_r) \tag{5}$$

知识蒸馏方法通过 KL 散度损失来匹配学生 $y_s = \text{Softmax}(a_s/\tau)$ 和教师 $y_t = \text{Softmax}(a_t/\tau)$ 的软化输出。其中,KL 散度也称为相对熵(Relative Entropy),是用来衡量两个概率分布之间差异的一种度量方式。其衡量的是当用一个分布 Q 来拟合真实分布 P 时所需要的额外信息的平均量,公式如下:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{6}$$

其中,x 代表概率分布中一个可能的事件或状态。$P(x)$ 和 $Q(x)$ 分别表示真实概率分布和模型预测的概率分布中事件 x 的概率。KL 散度通常用于无监督学习任务(如聚类、降维)。这类任务缺乏对应的标签信息,无法通过交叉熵评估模型性能,因此需要一种量化模型预测分布与真实分布之间差异的方法,KL 散度恰好可以实现这一目标,知识蒸馏中引入 KL 散度构建的损失项如下式所示:

$$L_{KD} = \tau^2 \text{KL}(y_s, y_t) \tag{7}$$

其中,y_s 为学生模型的输出分布,y_t 为教师模型的输出分布。引入了温度参数 τ 来对教师模型的输出结果的软化进行额外控制,之后学生模型使用如下损失函数进行训练:

$$L_{student} = (1 - \lambda) L_{SL} + \lambda L_{KD} \tag{8}$$

其中,λ 是控制两种损失之间权衡的超参数,L_{SL} 为学生模型的标准监督报告。本文将这种知识蒸馏方式称为基线知识蒸馏[16]。

本文提出基于助教的知识蒸馏增强神经机器翻译方法,通过引入中间规模的助教模型来缓解教师模型预测输出中知识丰富程度不足的问题。由于教师模型的翻译性能很高,因此其预测输出中正标签的预测概率很高,从而导致负标签中包含的不同预测结果间的相关信息等不够充足,减少了传递给学生模型的信息量,从而影响了知识蒸馏效率。引入助教后,这部分信息得到补充,知识蒸馏效率和学生模型性能得以提升。本文提出方法的训练框架如图 2 所示,其中 Soft Targets 的含义是指教师网络输出的概率分布经温度参数 τ 软化后的结果。这个软化的概率分布对每个输出类别都分配了概率,其中正标签的概率最高,分布比较平缓。在训练阶段,教师网络的 Logits 输出除以温度参数 τ 后再做 Softmax 变换,得到这个软化的概率分布。

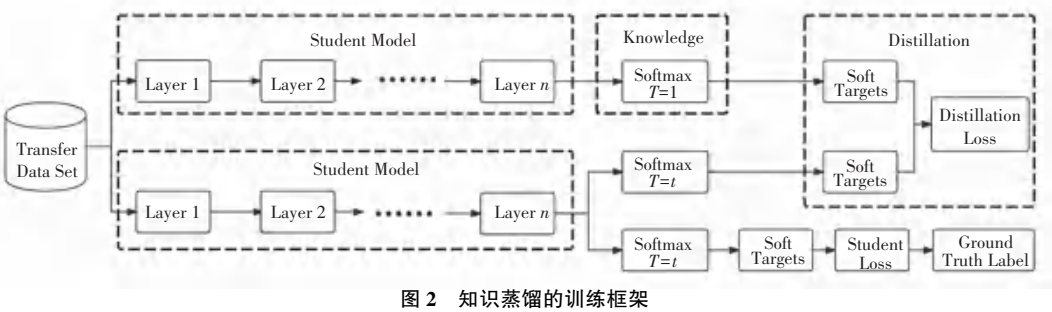

图 2 知识蒸馏的训练框架

Fig. 2 Training framework for knowledge distillation

4　实验结果与分析

4.1　数据预处理与模型设置

4.1.1　数据预处理

为了验证本文方法的有效性，选择了公开标准的 WMT14[17] 和 WMT16[18] 数据集进行实验。为了与已有的知识蒸馏方法进行公平对比，本文将 WMT14 英语-德语(En-De)、WMT14 英语-法语(En-Fr)和 WMT16 英语-罗马尼亚语(En-Ro)3 个翻译任务的数据设置与 Zhang[19] 等的研究保持一致。本文对这 3 个数据集分别进行了标点符号规范化、分词[20]、BPE(字节对编码)和二值化的预处理操作。对 WMT16 土耳其语-英语也采取同样的预处理操作确保实验结果具备可对比性。

4.1.2　模型设置

为了验证本文方法的有效性，本文采用了开源的 fairseq 工具包，并以 Transformer 作为基准模型。标准 Transformer 模型的结构包括六层编码器和解码器。在训练学生模型时，本文使用了标准的 Transformer 模型，设置学习率和 dropout 分别为 7e-4 和 0.1。前馈神经网络的维度为 2 048，注意力头的个数为 8，输入词向量维度为 512，训练步数 num_updates 设置为 200 000，其他参数保持默认。

对于教师模型，本文采用了 Transformer_vaswani_wmt_en_de_big 模型架构，输入词向量维度为 1 024，前馈神经网络的维度为 4 096，学习率设置为 5e-4，训练步数 num_updates 设置为 300 000。由于教师模型规模较大，本文将其训练时间延长，以确保教师模型具备足够的性能表现，从而使知识蒸馏的效果更加明显。对于助教模型，本文将其规模控制在学生模型和教师模型之间，因此将其前馈神经网络维度和输出词向量维度均设置为学生模型和教师模型的中间值。在模型评估方面，本文采用了不区分大小写的机器双语互译评估 BLEU[21] 得分来评估模型的翻译质量。BLEU 具体计算如下式：

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{9}$$

其中，p_n 是 n-gram 精确度，即匹配的 n-gram 数量除以总的 n-gram 数量；w_n 是对数平均权重，通常设置为(1/N)；BP 是惩罚因子(Brevity Penalty)，用来惩罚过短的翻译输出。对于每个翻译任务，本文保留了最后十个模型进行平均，并使用 multi-BLEU.perl3 脚本进行评测。

4.2　实验与分析

4.2.1　基于助教的知识蒸馏方法(ATKD)的性能

本文提出的基于助教的知识蒸馏增强神经机器翻译方法，在 WMT14 英语-德语(En-De)、英语-法语(En-Fr)、WMT16 英语-罗马尼亚语(En-Ro)和土耳其语-英语(Tr-En)4 个翻译任务上验证性能。实验结果见表 1。从实验数据可以看出，本文的方法在上述 4 个语言方向的翻译任务上均有提升，特别是在英语-罗马尼亚语翻译任务上 BLEU 得分提高了 0.72 分，展现出了良好的性能。

表 1　ATKD 在不同翻译任务实验中的 BLEU 分数对比

Table 1　Comparison of BLEU scores of ATKD in different translation tasks

方法	英语-德语	英语-法语	英语-罗马尼亚语	土耳其语-英语
基线	28.59	41.13	33.65	25.31
ATKD	29.13	41.65	34.37	25.62

4.2.2　与其他知识蒸馏方法在不同翻译任务上的比较

为了进一步验证本文提出的基于助教的知识蒸馏增强神经机器翻译方法(ATKD)的有效性，本文对 ATKD 和其他知识蒸馏方法进行了对比实验，并将实验结果汇总在表 2 中。其中，Student 代表知识蒸馏中的学生模型，Teacher 代表知识蒸馏中的教师模型。

表 2　在不同翻译任务上与其他知识蒸馏方法 BLEU 分数对比

Table 2　Comparison of BLEU scores with other knowledge distillation methods on different translation tasks

知识蒸馏方法	英语-德语 BLEU	英语-法语 BLEU	英语-罗马尼亚语 BLEU
Student($\mathrm{Transformer}_{\mathrm{base}}$)	27.42	40.97	33.59
+Word-KD[22]	28.03	41.10	33.77
+Seq-KD[22]	28.22	41.44	33.69
+Annealing KD[23]	27.91	41.20	33.67
+Seer Forcing[23]	27.56	40.97	33.77
本文方法	**29.13**	**41.65**	**34.21**
Teacher($\mathrm{Transformer}_{\mathrm{big}}$)	29.44	42.98	34.70

为了确保公平比较实验结果，本文在数据集的设置和数据来源上与文献[17]保持一致。通过对 WMT14 英语-德语、WMT14 英语-法语和 WMT16 英语-罗马尼亚语翻译任务的分析，可以看出引入助教模型后，由知识蒸馏得到的学生模型在上述 3 个语言方向翻译任务的 BLEU 分数均高于其他方法。进一步证实了本文提出方法的有效性和鲁棒性。

4.2.3 校验集 BLEU 和损失

为了评估基于助教的知识蒸馏增强神经机器翻译方法对学生模型训练过程中收敛情况的影响,验证本文提出的知识蒸馏框架在实验过程中的表现,本文在 WMT14 英语-德语翻译任务上进行了计算。本研究评估了模型训练过程中不同轮次上校验集的 BLEU 得分和损失情况,结果如图 3 所示。

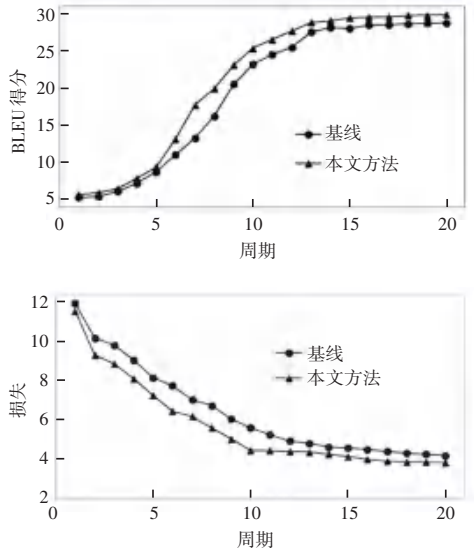

图 3 WMT14 英语-德语翻译任务不同轮次校验集 BLEU 得分和损失对比

Fig. 3 Comparison of validation BLEU scores and losses during training on the WMT14 EN-Pe task

从实验结果可以明显看出,相较于基线模型,本文提出的知识蒸馏框架具有更好的损失表现和更高的 BLEU 得分。这一效果的实现归功于助教模型的引入,其能够将负标签中包含的知识更充分地传递给学生模型,从而有效提高了学生模型的翻译性能。

5 结束语

本文从知识蒸馏的模型视角出发,提出了基于助教的知识蒸馏增强神经机器翻译方法,旨在解决师生模型在规模差距过大时知识蒸馏效率下降,导致学生模型性能不佳的问题。该方法旨在将大规模预训练神经机器翻译模型的性能更有效地迁移到小模型中,从而充分提升小模型的性能,特别适用于神经机器翻译领域。通过在 WMT14 和 WMT16 数据集上开展对比实验,结果表明,相较于基线系统,本文提出的方法进一步提升了神经机器翻译模型的性能表现。在未来的研究中,希望可以继续从 Transformer 模型复杂结构入手,探索知识蒸馏过程中教师模型与学生模型在结构上的联系,以进一步完善知识蒸馏方法在神经机器翻译领域的应用。

参考文献

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

[2] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[3] JIN Chang , DUAN Renchong, XIAO Nini, et al. Inter-layer knowledge distillation for neural machine translation [C]// Proceedings of the 20th Chinese National Conference on Computational Linguistics. CCL, 2021: 166-175.

[4] GUMMA V, DABRE R, KUMAR P. An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models[J]. arXiv preprint arXiv,2304.09388, 2023.

[5] ZHOU Xiaoqing, DUAN Xiangyu, YU Hongfei, et al. Progressive semi - knowledge distillation for neural machine translation[J]. Journal of Chinese Information Processing, 2021, 35(2): 52-60.

[6] WANG F, YAN J, MENG F, et al. Selective knowledge distillation for neural machine translation [J]. arXiv preprint arXiv, 2105.12967, 2021.

[7] LIU W, ZHOU P, ZHAO Z, et al. Fastbert: A self-distilling bert with adaptive inference time [J]. arXiv preprint arXiv, 2004.02178, 2020.

[8] JIANG Y, WANG X, BACH N, et al. Structure-level knowledge distillation for multilingual sequence labeling[J]. arXiv preprint arXiv, 2004.03846, 2020.

[9] ZHANG C, LIU J, DANG K, et al. Multi-scale distillation from multiple graph neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI,2022: 4-12.

[10] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020: 5191-5198.

[11] JAFARI A, REZAGHOLIZADEH M, SHARMA P, et al. Annealing knowledge distillation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. EACL, 2021: 2493-2504.

[12] 孙潇,朱聪慧,赵铁军. 融合翻译知识的机器翻译质量估计算法[J]. 智能计算机与应用,2019,9(2):271-274.

[13] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv,1409.0473.2014.

[14] 孙海鹏,赵铁军. 无监督神经机器翻译综述[J]. 智能计算机与应用, 2021,11(2):1-6.

[15] OTT M, EDUNOV S, BAEVSKI A, et al. Fairseq: A fast, extensible toolkit for sequence modeling[J]. arXiv preprint arXiv, 1904.01038,2019.

[16] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv, 1503.02531,2015.

[17] TAKASE S , KIYONO S. Rethinking perturbations in encoder-decoders for fast training [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for

Computational Linguistics. NAACL，2021：5767-5780.

[18] SHEN Y，CHU C H，CROMIERES F，et al. Cross-language projection of dependency trees with constrained partial parsing for tree-to-tree machine translation[C]//Proceedings of the First Conference on Machine Translation. WMT，2016：1-11.

[19] ZHANG C，LIU J，DANG K，et al. Multi-scale distillation from multiple graph neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI，2022：4337-4344.

[20] ZHANG S，LIANG Y，WANG S，et al. Towards understanding and improving knowledge distillation for neural machine translation[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics. ACL，2023：8062-8079.

[21] PAPINENI K，ROUKOS S，WARD T，et al. Bleu：A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. ACL，2002：311-318.

[22] KIM Y，RUSH A M. Sequence-level knowledge distillation[J]. arXiv preprint arXiv，1606.07947，2016.

[23] FENG Y，GU S H，GUO D J，et al. Guiding teacher forcing with seer forcing 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. IJCNLP，2021：2862-2872.