文章编号：1003-0077(2019)07-0046-10

利用单语数据改进神经机器翻译压缩模型的翻译质量

李 响[1,2]，刘 洋[3]，陈 伟[4]，刘 群[5]

（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；
3. 清华大学 计算机科学与技术系，北京 100084；
4. 北京搜狗科技发展有限公司 语音交互技术中心，北京 100084；
5. 华为 诺亚方舟实验室，香港）

摘 要：该文提出利用一个大型且精度高的神经机器翻译模型（教师模型）从单语数据中提取隐性双语知识，从而改进小型且精度低的神经机器翻译模型（学生模型）的翻译质量。该文首先提出了“伪双语数据”的教学方法，利用教师模型翻译单语数据获得的合成双语数据改进学生模型，然后提出了“负对数似然—知识蒸馏联合优化”教学方法，除了利用合成双语数据，还利用教师模型获得的目标语言词语概率分布作为知识，从而在知识蒸馏框架下提高学生模型的翻译质量。实验证明，在中英和德英翻译任务上，使用该方法训练的学生模型不仅在领域内测试集上显著超过了基线学生模型，而且在领域外测试集上的泛化性能也得到了提高。

关键词：神经机器翻译；知识蒸馏；单语数据

中图分类号：TP391 **文献标识码**：A

Improving the Translation Quality of Compressed Neural Machine Translation Models with Monolingual Data

LI Xiang[1,2], LIU Yang[3], CHEN Wei[4], LIU Qun[5]

(1. Key Larboratory 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;
3. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
4. Voice Interaction Technology Center, Sogou Inc., Beijing 100084, China;
5. Huawei Noah's Ark Lab, Hong Kong, China)

Abstract: This paper proposes to utilize a large and high precision neural machine translation (NMT) model (teacher model) to distill invisible bilingual knowledge from monolingual data in order to improve the translation quality of a small and low precision NMT model (student model). This paper first proposes the method of pseudo bilingual data where the student model is improved based on the synthesized training data by utilizing the teacher model to translate the monolingual data. Further, this paper proposes the joint optimization approach of negative log-likelihood and knowledge distillation. In addition to the synthetic training data, the student model can be enhanced by using the probability distribution of target language words obtained by the teacher model as knowledge under the knowledge distillation framework. Experiments on the Chinese-English and Germany-English translation tasks show that the student model trained by the proposed approaches not only significantly outperforms the baseline student model regarding translation quality on in-domain test sets, but also achieves a better generalization performance on an out-domain test set.

Keywords: neural machine translation; knowledge distillation; monolingual data

收稿日期：2018-09-17 **定稿日期**：2018-11-06

基金项目：国家自然科学基金(61876174,61662077)

0　引言

近年来，端到端神经机器翻译[1-3]（neural machine translation，NMT）作为一种崭新的机器翻译方法，在多个语言对上都获得了目前最佳的翻译质量[4]，显著优于统计机器翻译[5-6]，成为当前主流的机器翻译方法。

NMT 通常需要庞大复杂的网络结构才能获得高质量译文，例如，Vaswani 等[7]开发的一个包含数亿参数的深层 NMT 模型在 WMT 2014 英德和英法翻译任务上取得优胜。然而，由于移动设备一般只配备计算能力较低的图形处理器（GPU）和有限的存储，因此难以在这些低资源移动设备上直接部署和运行复杂的 NMT 模型。随着移动端机器翻译需求的快速增加，对 NMT 模型的有效压缩成为一项重要的基础工作。

研究人员提出了多种用于压缩复杂神经网络的方法[8]，主要包括参数量化[9]、权重剪枝[10]和低秩分解[11]等。除此以外，作为一种可以兼顾压缩模型和改善模型预测精度的有效方法，知识蒸馏[12]（knowledge distillation，KD）已经被广泛应用于目标检测[13]等多个领域。知识蒸馏的目标是将预训练好的复杂模型输出的“知识”作为监督信号来训练另外一个网络参数规模更小的简单模型。这个复杂网络称为教师模型，而简单网络称为学生模型。知识蒸馏方法也被应用于改进压缩后 NMT 模型的翻译质量[14]，然而，用于指导学生模型的“知识”仅来源于有限的双语平行语料，这一方面限制了学生模型获取“知识”的边界，另一方面也限制了教师模型的潜力。因此，进一步改进压缩模型的翻译精度对于增强移动端离线机器翻译的实用性具有重要意义。

基于知识蒸馏，本文提出了一种简单而有效的利用大规模单语数据进一步改进学生 NMT 模型的方法。基本思路是利用教师模型从大规模单语数据中提取隐性双语“知识”，从而指导学生模型学习有限的双语数据中缺失的翻译映射关系。本文从不同维度提出了“教师模型生成伪双语数据”和“负对数似然—知识蒸馏联合优化”两种教学方法。在“教师模型生成伪双语数据”教学方法中，教师模型通过翻译源语言单语数据产生“伪”双语数据，由于教师模型一般具有比学生模型更强的预测能力，这些包含教师模型“知识”的“伪”双语数据既可以单独作为训练数据，也可以和真实双语语料合并得到增强的训练数据，并用于学生模型的训练，而无须改变学生模型原有的训练方法。在“负对数似然—知识蒸馏联合优化”教学方法中，除了利用“伪”双语数据，还通过修改学生模型的优化目标，在原有的 NMT 负对数似然（negative log-likelihood，NLL）优化目标基础上，同时引入知识蒸馏优化目标，从而将教师模型的词语预测能力更好地迁移到学生模型。在 NIST 汉英翻译和 WMT 德英翻译任务上的实验结果表明，利用本文提出的方法训练学生模型，不仅在领域内测试集上显著优于基线学生模型和只采用真实双语数据的学生模型，而且在领域外测试集上的泛化性能也得到了提高。

1　研究背景

1.1　神经机器翻译

图 1 表示基于循环神经网络编码器—解码器的 NMT 基本框架，由于循环神经网络具有很强的时序建模能力，编码器采用一个连续空间中的源语言语义向量 $\boldsymbol{c}$ 来表示一个任意长度的源语言词语序列，而解码器则对 $\boldsymbol{c}$ 进行还原，并生成目标语言词语序列。

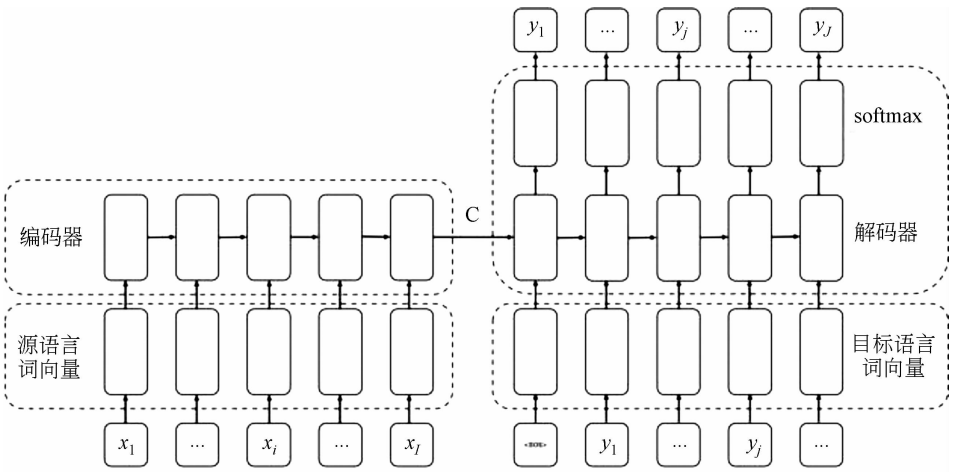

图 1　基于编码器-解码器的神经机器翻译模型

用 $\boldsymbol{x}=x_1,\cdots,x_i,\cdots,x_I$ 和 $\boldsymbol{y}=y_1,\cdots,y_j,\cdots,y_J$ 分别表示源语言和目标语言序列，NMT 通常采用式(1)对翻译概率建模：

$$P(\boldsymbol{y}\mid\boldsymbol{x};\boldsymbol{\theta})=\prod_{j=1}^{J}P(\boldsymbol{y}_j\mid\boldsymbol{x},\boldsymbol{y}_{<j};\boldsymbol{\theta}),\quad(1)$$

其中，$\boldsymbol{\theta}$ 表示模型参数，$\boldsymbol{y}_{<j}=y_1,\cdots,y_{j-1}$ 表示已经完成翻译的部分目标语言文本，而解码器采用式(2)预测第 j 个目标语言单词：

$$P(\boldsymbol{y}_j\mid\boldsymbol{x},\boldsymbol{y}_{<j};\boldsymbol{\theta})=\mathrm{softmax}(g(\boldsymbol{h}_j,\boldsymbol{y}_{j-1},\boldsymbol{c})),(2)$$

其中，$\boldsymbol{h}_j$ 表示第 j 个时刻解码器的隐层状态，$\boldsymbol{y}_{j-1}$ 表示第 $j-1$ 个目标语言词语的词向量，$\boldsymbol{c}$ 表示整个源语言句子的语义表示，通过激活函数 $g(\cdot)$ 进行非线性变换，最后采用 softmax 函数获得该时刻目标语言词汇的概率分布。

给定双语平行语料 $\mathcal{D}_{x,y}=\{<x^m,y^m>\}_{m=1}^{M}$，通常的 NMT 训练是以 NLL 函数作为损失函数，搜索最优参数 $\hat{\boldsymbol{\theta}}$ 使损失最小，如式(3)所示。

$$\hat{\boldsymbol{\theta}}=\arg\min_{\theta}\{\mathcal{L}_{\mathrm{NLL}}(\theta,\mathcal{D}_{x,y})\},\quad(3)$$

其中，NLL 的定义如式(4)所示。

$$\mathcal{L}_{\mathrm{NLL}}(\boldsymbol{\theta},D_{x,y})\quad(4)$$

$$=-\sum_{m=1}^{M}\sum_{j=1}^{J^m}\sum_{y\in Y}\mathbb{1}\{\boldsymbol{y}_j^m=\boldsymbol{y}\}\times\log P(\boldsymbol{y}\mid\boldsymbol{x}^m,\boldsymbol{y}_{<j}^m;\boldsymbol{\theta})$$

其中，Y 是目标语言词汇表，J^m 是第 m 个目标语言序列的长度，$\mathbb{1}\{\cdot\}$ 是指示函数。

式(3)表明，模型参数直接影响了神经机器翻译精度。一般来说，更多的参数有助于学习更复杂的双语映射关系。因此，采用复杂的神经机器翻译网络结构成为一种趋势，然而这也极大地增加了在低资源移动终端上使用 NMT 的计算和存储开销。

1.2 知识蒸馏

在神经机器翻译等分类任务中，一般采用一位有效编码(one-hot)w_i 标注的目标语言词语作为预测目标，如式(5)所示。

$$w_i\in\{0,1\}^{|y|},i=1,2,\cdots,T\quad(5)$$

其中，w_i 的维度与目标语言词汇表大小 $|Y|$ 相同，并且只有一个维度上有值为 1，这个位置对应目标单词在词汇表 Y 的位置，而其余值全为 0。因此，式(4)中的指示函数 $\mathbb{1}\{\cdot\}$ 表明只有对当前目标语言词语的预测概率才影响模型的优化，因此，one-hot 编码丢失了目标语言词汇表中的词间相似性信息。例如，单词“USA”“US”和“America”是同义词，但是使用上述方法，完全没有考虑到这种相似性。对于已经训练好的模型，当预测单词“USA”时，目标端词语预测概率很可能是：“USA”为 0.9，“US”为 0.05，“America”为 0.05。

为了弥补学生模型优化中分类监督信号不足的问题，Hinton 等[12]通过引入“温度”T 调整 softmax 函数的输入，如式(6)所示，从而提出了知识蒸馏学习方法，如图 2 所示。

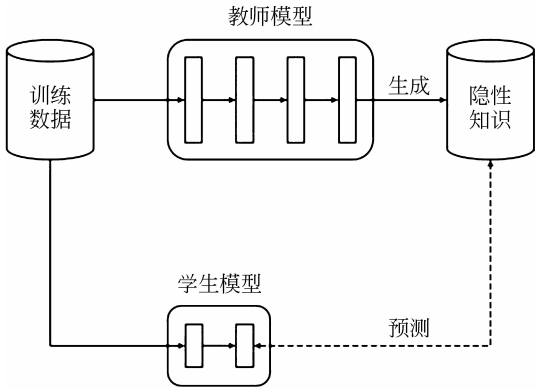

图 2 知识蒸馏学习方法

$$q_i=\frac{\exp(z_i/T)}{\sum_{j}^{|Y|}\exp(z_j/T)},\quad(6)$$

其中，当 $T=1$ 时，就变成标准的 softmax 函数，而当 $T>1$ 时，softmax 函数的输出被软化，每个预测类别间的概率差距就会变小。相比于 one-hot 编码的硬标签(ground truth label)，对教师模型采用这种方法，可以得到含有更多类间信息的软标签(soft target label)。

在学生模型训练中，将教师模型产生的软标签作为预测对象，则教师模型可以用于指导学生模型的训练，传统的 NLL 损失函数可以替换为式(7)所示的知识蒸馏损失函数 KD，让学生模型的词语概率分布 P 尽量拟合教师模型的词语概率分布 Q。

$$\mathcal{L}_{\mathrm{KD}}(\boldsymbol{\theta},\boldsymbol{D}_{x,y})\quad(7)$$

$$=-\sum_{m=1}^{M}\sum_{j=1}^{J^m}\sum_{y\in Y}\boldsymbol{Q}\times\log P(\boldsymbol{y}\mid\boldsymbol{x}^m,\boldsymbol{y}_{<j}^m;\boldsymbol{\theta})$$

2 单语增强的神经机器翻译压缩模型

知识蒸馏已经用于神经机器翻译模型的压缩和翻译质量的改进[14]。基本思想是首先根据双语数据训练一个教师模型，如式(8)所示。

$$\hat{\theta}_T=\arg\min_{\theta_T}\{L_{\mathrm{NLL}}(\theta_T,\mathcal{D}_{x,y})\},\quad(8)$$

其中，$\hat{\theta}_T$ 表示教师模型的参数。

最后利用训练后的教师模型指导学生模型的训练，如式(9)所示。

$$\mathcal{L}_{\mathrm{KD}}(\boldsymbol{\theta}_{\mathrm{S}},\hat{\boldsymbol{\theta}}_{T},\boldsymbol{D}_{x,y})=-\sum_{m=1}^{M}\sum_{j=1}^{J^{m}}\sum_{y\in y}\boldsymbol{Q}(y\mid x^{m},y_{<j}^{m},\hat{\theta}_{T})\times\log P(\boldsymbol{y}\mid\boldsymbol{x}^{m},\boldsymbol{y}_{<j}^{m};\boldsymbol{\theta}_{\mathrm{S}}) \quad (9)$$

然而，高质量的双语数据极为有限，制作成本极高，尤其在低资源机器翻译任务中更为明显，制约了学生模型获取翻译“知识”的来源和翻译精度的进一步改进。因此，本文提出了“教师模型生成伪双语数据”和“负对数似然—知识蒸馏联合优化”两种教学方法，利用教师模型从大规模源语言单语数据中提取隐性双语“知识”，从而改进学生模型的翻译质量和泛化能力。

2.1　教师模型生成伪双语数据教学

式(3)表明，除了模型参数，训练数据对于模型预测能力也有重要影响。因此，本文首先从数据增强的角度提出了“教师模型生成伪双语数据”教学方法(PT)，在无需改变学生模型原有的训练方式的条件下，改进学生模型的翻译精度。具体步骤如下：

① 首先用真实双语平行语料 $D_{x,y}$ 训练一个教师模型 $\hat{\theta}_{T}$；

② 通过束搜索(beam search)解码，利用教师模型对一个大规模源语言单语数据 $D_{\widetilde{x}}$ 进行翻译，获得“伪”双语数据 $D_{\widetilde{x},\widetilde{y}}$；

③ 合并真实双语数据和“伪”双语数据，从而得到一个更大规模的平行语料 $\widetilde{D}_{x,y}$：

$$\widetilde{D}_{x,y}=D_{x,y}\cup D_{\widetilde{x},\widetilde{y}};$$

④ 根据式(10)所示，采用 NLL 作为学生模型的损失函数：

$$\hat{\theta}_{\mathrm{S}}=\arg\min_{\theta_{\mathrm{S}}}\{L_{\mathrm{PT}}\}, \quad (10)$$

$$\mathcal{L}_{\mathrm{PT}}=L_{\mathrm{NLL}}(\theta_{\mathrm{S}},\widetilde{D}_{x,y}),$$

其中，θ_{S} 表示学生模型的参数。

图 3 表示基于训练数据 $\widetilde{D}_{x,y}$ 的 PT 方法。其中，教师模型包含两个隐藏层，而学生模型只包含一个隐藏层，并且隐藏层的神经元数量也更少。利用这个教师模型对单语数据解码，并选择损失最小的候选翻译作为最终译文，于是学生模型可以基于这些合成的平行语料进行训练。

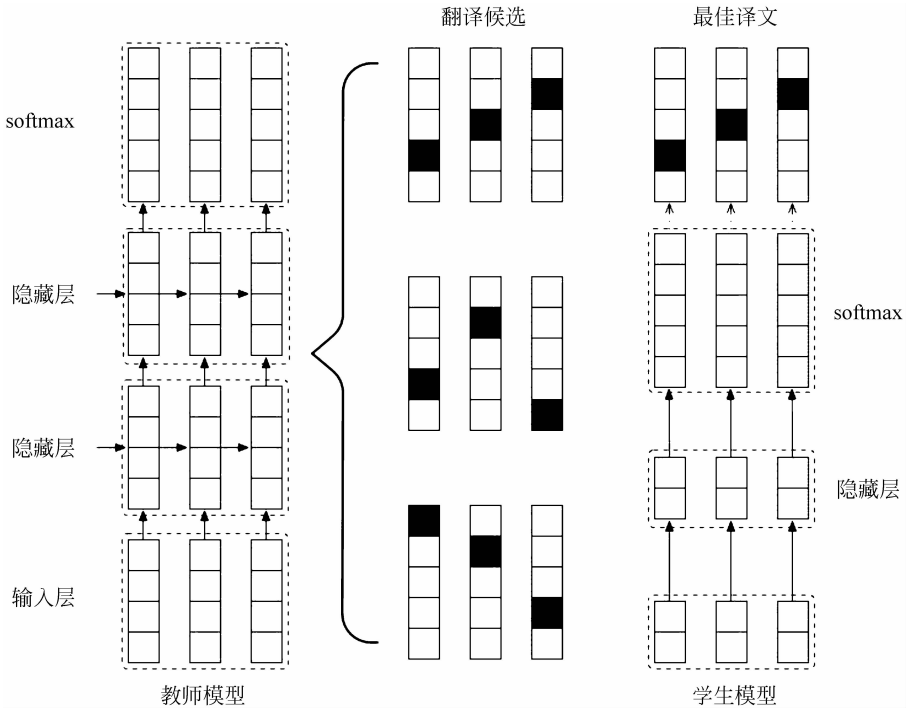

图 3　PT 方法示意图

本文提出的 PT 方法类似于机器学习中的“自训练”(self-training)方法，主要区别在于 PT 方法利用一个预训练好的教师模型生成“伪”平行数据，从大量未标注单语数据中提取双语翻译映射“知识”，用于指导学生模型的训练，而“自训练”方法一般是利用同一个模型迭代地生成“伪”数据，从而有效而低成本地利用未标注数据提升同一个模型的性能，其目的并不是用于改进压缩模型的预测精度。

2.2 负对数似然—知识蒸馏联合优化教学

神经机器翻译一般使用的束搜索解码算法存在几个明显的问题，包括曝光偏差(exposure bias)、损失评估失配(loss-evaluation mismatch)和标签偏差(label bias)等[15]，这表明教师模型通过束搜索产生的“伪”平行数据也可能存在错误或者噪声。

对于 PT 方法，教师模型仅仅提供句级的“伪”平行训练数据作为“知识”训练学生模型，尽管这种数据包含大量有价值的双语翻译知识，然而采用 NLL 损失函数的学生模型可能无法衡量“伪”双语数据的质量。因此，为了增强学生模型在合成数据上的学习能力，本文提出了“负对数似然—知识蒸馏联合优化”教学方法(JT)，除了 NLL 损失函数外，JT 教学还额外引入了 KD 损失函数，利用教师模型提供的软标签“知识”指导学生模型训练，步骤如下：

① 采用和 PT 教学同样的增广训练数据 $\widetilde{D}_{x,y}$；

② 根据式(11)优化学生模型：

$$\hat{\theta}_s = \arg\min_{\theta_S}\{\mathcal{L}_{\mathrm{JT}}\},$$

$$\mathcal{L}_{\mathrm{JT}} = \alpha\,\mathcal{L}_{\mathrm{NLL}}(\boldsymbol{\theta}_s, \hat{\boldsymbol{D}}_{x,y}) + (1-\alpha)\,\mathcal{L}_{\mathrm{KD}}(\boldsymbol{\theta}_s, \hat{\theta}_t, \hat{\boldsymbol{D}}_{x,y}) \tag{11}$$

其中，α 是用于平衡 NLL 和 KD 优化目标的超参数。

相比只采用 NLL 作为损失函数的 PT 方法，当额外引入 KD 损失函数时，对于学生模型将要预测的目标语言词语，教师模型也同步生成了对应的词汇表概率分布。因此，这些包含词间语义关系的软标签“知识”可以为学生模型的优化提供额外的指导信息，帮助学生模型更好地应对合成训练数据中的噪声。

图 4 为 JT 方法示意图。对于增广数据 $\widetilde{D}_{x,y}$ 中的一条双语数据，教师模型在目标端的每个单词位置产生概率分布，其中每个实数值表示该位置对应词表中单词的预测概率。因此，学生模型的训练需要同时对数据中的硬标签和教师模型产生的软标签进行联合优化。

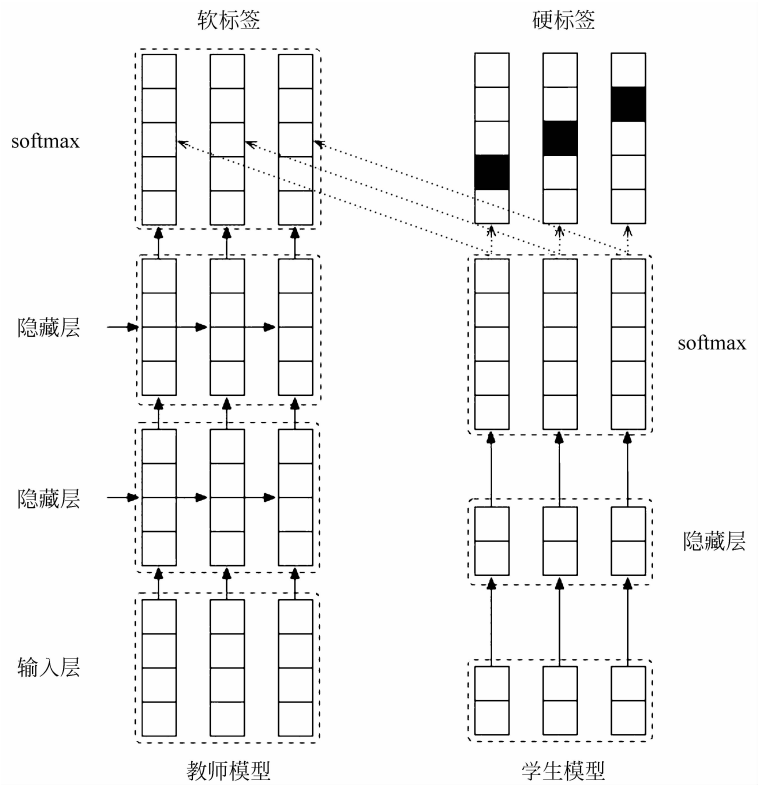

图 4 JT 方法示意图

通过联合使用 NLL 和 KD 优化函数，并借助高质量的增广训练数据，学生模型可以从两个不同的优化目标中学习多样的翻译知识，缓解了学生模型由于压缩网络结构导致泛化能力下降的问题。

3　实验

3.1　数据处理

我们分别在 NIST 汉英和 WMT 德英两个翻译任务上评价了本文提出的方法。

对于汉英任务，本文采用的训练数据包含来源于 LDC 语料库的 125 万汉英句对①，包括 2 790 万中文词语和 3 450 万英语词语。我们选择 NIST 2002（MT02）作为验证集，NIST 2003（MT03）、NIST 2004（MT04）、NIST 2005（MT05）和 NIST 2006（MT06）作为测试集②。我们从中国新闻网③采集了 2011 年期间的 300 万新闻领域汉语句子作为源语言单语数据。我们限制汉语和英语的词表只包含数据中最高频的 3 万个词，分别覆盖了 97.7% 和99.3%的训练数据。另外，只采用源端句子长度不超过 80 个词的训练数据。

对于德英任务，我们采用已经预处理后的包含 585 万句对的 WMT 2017 训练数据④，分别采用 newsdev2017 和newstest2017 作为验证集和测试集。我们从爱丁堡大学提供的 WMT 2016 英德反向翻译数据的德语部分随机选择 300 万句子作为源语言单语数据⑤。为了减少训练数据中的集外词，我们采用字节对编码（byte pair encoding，BPE）[16] 方法处理数据，我们限制源语言和目标语言词汇表都只包含词频最高的 3.7 万的子词。

另外，对于汉英和德英翻译任务，都采用一个特殊的符号“＜UNK＞”表示数据中的集外词。

3.2　实验设置

本文实验中所有的系统都是基于神经机器翻译工具包 Nematus⑥ 实现的。需要指出的是，本文提出的方法主要是改变学生模型的数据来源或优化目标，与具体的网络结构无关，因此适用于包括 Transformer[7] 在内的任何神经机器翻译模型。

本文采用 Nematus 默认的单层网络结构，每个 batch 最多包含 2 048 个源语言单词，采用 dropout[17] 缓解模型过拟合问题，源语言和目标语言单词丢弃率设置为 0.1，而神经元的丢弃率（dropout）设置为 0.2，采用 Adam[18] 方法优化模型参数，学习率设置为 0.1，采用层正则化方法[19] 加速模型收敛和提高训练稳定性，同时，每 5 000 个 batch 对验证集做一次解码，如果连续 10 次验证集译文的 BLEU 得分都没有提高，则提前终止模型训练，避免模型过拟合。由于汉英训练数据规模较小，为了避免参数初始化导致的测试误差，我们对所有的汉英系统均训练两个不同的模型，每个模型采用不同的随机种子，最后取两个模型的均值作为最终实验结果。

在实验中，对开发集和验证集进行解码时，我们将束搜索大小设置为 10。为了产生质量较高的“伪”平行数据，同时也兼顾解码效率，在使用教师模型对单语数据生成译文时，束搜索大小设定为 4。

我们采用大小写不敏感的 NIST BLEU 作为评价指标。

3.3　模型设置

我们采用的神经机器翻译系统如下。

① 教师模型（Teacher）：词向量和隐层神经元个数都设置为 512，共包含 5 640 万个参数；

② 学生模型（Student）：词向量和隐层神经元个数都设置为 128，共包含 1 220 万个参数，相比教师模型减少了 78%，模型得到了显著压缩。在基线学生模型基础上，分别实现了 PT 学生模型和 JT 学生模型。对于 JT 学生模型，我们对汉英任务分别设置 α 为[0.3，0.4，0.5，0.6，0.7]，发现 0.5 时在开发集上的性能最好，因此我们在汉英和德英的 JT 学生模型中都采用 0.5 作为 α 的默认配置。

我们也与 Kim 和 Rush[14] 提出的三种模型进行了比较，这些方法在训练过程中仅使用双语数据作为指导学生模型的知识来源：

① Word-KD：采用 KD 优化目标，直接利用教师模型提供的软标签指导学生模型的训练；

② Seq-KD：教师模型对双语训练数据的源端文本生成候选译文，并根据目标端参考译文选择最相近的作为最终译文，得到“伪”训练数据，通过合并原始双语数据和合成训练数据，采用 NLL 优化目标训练学生模型；

③ Seq-KD＋Seq-Inter＋Word-KD（SSW）：首

---

① 语料包括 LDC2002E18，LDC2003E07，LDC2003E14，LDC2004T07 Hansards 部分，LDC2004T08 和 LDC2005T06。

② http://www.itl.nist.gov/iad/mig/tests/mt/

③ www.chinanews.com

④ http://data.statmt.org/wmt17/translation-task/preprocessed/de-en/

⑤ http://data.statmt.org/rsennrich/wmt16_backtranslations/en-de/news.bt.en-de.de.gz

⑥ https://github.com/EdinburghNLP/nematus

先仅在合成训练数据上训练学生模型，然后和原始训练数据合并，并采用NLL和KD联合优化目标指导学生模型的训练。

3.4 实验结果

3.4.1 Dropout对学生模型的影响

作为缓解深度神经网络过拟合问题的一种有效方法，dropout得到广泛应用。知识蒸馏也可以看作是一种模型正则化方法，因此，dropout是否会影响利用教师模型对学生模型的改进？为了验证这个猜测，我们在汉英和德英任务上，分别将全部单语数据作为知识源，并采用JT模型对这个问题进行了验证。

表1的实验结果表明，当采用dropout训练基线学生模型时，在汉英和德英任务上分别下降2.25 BLEU和1.75 BLEU，说明对于经过大幅压缩的学生模型，采用dropout显著损害了基线学生模型的预测能力，而对于教师模型来说，dropout则显著改进了模型的翻译质量；另一方面，不采用dropout的JT学生模型显著提升了基线学生模型的翻译质量，在汉英和德英任务上分别提升到到4.3 BLEU和3.25 BLEU，说明dropout对知识蒸馏产生了抑制作用，导致学生模型无法得到有效改进。因此，在后续实验中，我们只对教师模型的训练使用dropout。

表1 dropout对模型翻译质量的影响

模型	MT03	MT04	MT05	MT06	平均	newstest2017
Teacher(+dropout)	38.91	41.14	39.42	38.61	39.52	29.88
Teacher(−dropout)	35.52	38.05	36.17	35.19	36.23	28.39
Student(+dropout)	29.87	31.82	29.47	29.24	30.10	22.95
Student(−dropout)	32.04	34.58	31.44	31.34	32.35	24.70
JT(+dropout)	31.88	34.36	31.15	31.09	32.17	24.89
JT(−dropout)	36.68	38.35	35.49	36.14	36.72	27.95

我们认为神经网络产生过拟合的主要原因在于模型复杂度过高，导致模型容易陷入局部最优。而dropout通过引入随机性的神经元失活，限制了搜索空间，因此可以缓解模型陷入局部最优的问题。而知识蒸馏实际上是一种模型压缩方法，压缩后的模型产生过拟合的可能性下降，在这种情况下再采用dropout来限制搜索空间反而会导致欠拟合，降低了压缩模型的预测能力。

3.4.2 单语数据对于学生模型的影响

我们评价了使用单语数据作为额外知识来源对改进学生模型的作用，以及PT模型和JT模型的差别。我们采用全部单语数据作为知识蒸馏来源，实验如表2所示，实验表明：

表2 单语数据对学生模型的影响

模型	MT03	MT04	MT05	MT06	平均	newstest 2017
Teacher	38.91	41.14	39.42	38.61	39.52	29.88
Student	32.04	34.58	31.44	31.34	32.35	24.70
Word-KD	34.97	37.47	33.89	34.46	35.20	25.63
Seq-KD	35.66	37.72	34.27	35.83	35.87	26.27
SSW	35.99	37.85	34.68	36.04	36.14	26.62
PT	36.16	37.69	34.89	35.36	36.03	27.38
JT	36.68	38.35	35.49	36.14	36.72	27.95

① 本文提出的PT和JT方法都显著改进了基于双语训练数据的基线学生模型的翻译精度，在汉英和德英任务上的提升幅度分别达到了4.37 BLEU和3.25 BLEU；

② 本文提出的 JT 方法较 PT 方法在汉英和德英任务上的最大提升达到了 0.69 BLEU 和 0.57 BLEU，说明在基于 NLL 的训练中融入 KD 优化目标可以有效提升学生模型的翻译精度；

③ 在汉英和德英任务上，本文提出的 JT 模型相比 Kim 和 Rush[14] 提出的 SSW 模型分别提高了 0.58 BLEU 和 1.33 BLEU，说明仅利用双语训练数据作为知识来源限制了教师模型预测能力的发挥，同时也限制了学生模型学习外部知识的机会，而大规模的单语数据可以缓解这些限制，进一步改进学生模型。

3.4.3　单语数据规模对学生模型的影响

由于单语数据量远远大于双语数据，我们随机选取不同数量的单语数据，评估单语数据规模对学生模型的影响。

表 3　单语数据规模对学生模型的影响

模型	单语规模	MT03	MT04	MT05	MT06	平均	newstest 2017
Teacher	—	38.91	41.14	39.42	38.61	39.52	29.88
Student	—	32.04	34.58	31.44	31.34	32.35	24.70
PT	1M	35.63	37.24	34.40	34.80	35.52	26.53
	2M	35.85	37.41	34.79	35.17	35.81	27.19
	3M	36.16	37.69	34.89	35.36	36.03	27.38
JT	1M	35.98	37.75	34.88	35.32	35.98	26.81
	2M	36.39	38.02	35.22	35.78	36.35	27.51
	3M	36.68	38.35	35.49	36.14	36.72	27.95

表 3 的实验结果显示，当单语数据规模从 100 万句增加到 300 万句时，汉英和德英任务上的 JT 模型翻译质量分别提高了 0.74 BLEU 和 1.14 BLEU，说明增加单语数据有助于改进学生模型的翻译质量。同时，实验结果也表明，在同样规模的单语数据情况下，汉英和德英任务上的 JT 方法都要优于 PT 方法。

3.4.4　缺少双语训练数据对学生模型的影响

实际当中，企业出于数据安全的考虑，权限较低的工程师或者第三方开发者往往只能够获得教师模型，而无法直接获得原始双语训练数据，导致无法基于真实双语数据改进学生模型。在这种情况下，我们在汉英和德英任务上分别只采用教师模型生成的 300 万“伪”双语数据，并采用 JT 模型探索只有单语数据对学生模型的影响。我们用 OnlyMono-JT 表示只使用“伪”双语数据的 JT 学生模型。表 4 的实验结果表明，在缺乏真实的双语数据时，在仅依靠单语作为知识源的汉英和德英任务上，JT 学生模型翻译精度的提升分别达到了 4.03 BLEU 和 2.69 BLEU，说明单语数据的利用显著改进了学生模型的翻译质量。同时，结果也表明，在知识蒸馏训练中引入原始有标注的训练数据，能够有效地提升学生模型的精度。

表 4　缺乏双语平行数据对学生模型的影响

模型	MT03	MT04	MT05	MT06	平均	newstest 2017
Teacher	38.91	41.14	39.42	38.61	39.52	29.88
Student	32.04	34.58	31.44	31.34	32.35	24.70
JT	36.68	38.35	35.49	36.14	36.72	27.95
OnlyMono-JT	36.38	38.13	35.07	35.89	36.38	27.39

3.4.5　领域外数据的泛化能力

由于实验采用的训练数据和单语数据来源于新闻领域，为了评价本文提出的方法在领域外数据上的泛化性能，我们采用联合国平行语料 1.0 版（United Nations Parallel Corpus v1.0）[20] 的汉英测试数据集（UM-testing），涉及多个不同的领域，包括新闻、口语、法律、教育、科技和字幕。实验系统为使用 300 万单语数据的汉英词级模型。

表 5 的实验结果表明，本文提出的方法在领域

外测试集上具有最好的泛化能力，说明单语数据可以有效改进学生模型的泛化能力。

表5 JT学生模型在领域外的联合国语料汉英测试集上的翻译质量

模型	UM-testing
Teacher	17.15
Student	11.89
SSW	13.68
JT	14.74

4 相关工作

知识蒸馏已经得到广泛应用。Buciluǎ等[21]证明可以将一个集成模型的知识压缩在一个单独的较小模型中，从而更容易部署。Ba等[22]利用一种相对教师模型更浅的学生模型来模仿深层神经网络，并保证两者的参数量相同，因此学生模型的每一层也更宽，实验结果表明浅层网络也可以表现出接近深层网络的性能。Chen等[23]提出用教师—学生模型框架改善低资源的神经机器翻译。Anil等[24]提出在线蒸馏方法，提高了大规模分布式神经网络训练的模型精度，并加快了训练速度。

神经机器翻译主要利用有限的双语数据训练模型，因此，研究人员提出了采用单语数据改进翻译性能的方法。Zhang和Zong[25]提出借助源语言单语数据改进翻译质量，而Sennrich等[26]提出借助目标语言单语数据的反向翻译方法改进了神经机器翻译质量。Cheng等[27]提出通过一种半监督的方式，采用源端—目标端和目标端—源端两个翻译模型分别作为编码器和解码器，并在训练过程中对源语言和目标语言单语数据进行重构，通过联合训练方式有效利用单语数据，在汉英测试集上得到显著改进。Domhan和Hieber[28]提出结合神经机器翻译解码器和目标语言模型，利用大规模目标语言单语数据显著改进稀缺语之间的翻译精度。

5 总结与展望

在知识蒸馏框架下，本文提出了利用大规模源语言单语数据改进学生神经机器翻译模型的有效方法。我们分别提出了“教师模型生成伪双语数据”和“负对数似然—知识蒸馏联合优化”两种学生模型教学方法，从而使教师模型可以从不同的维度利用单语数据指导学生模型的训练。基本思想是单语数据可以为学生模型提供更多双语数据无法提供的翻译知识，从而增强学生模型在真实条件下的泛化能力。在NIST汉英和WMT德英数据集上，实验结果表明本文提出的方法极大地改进了基线学生模型和其他仅使用双语数据作为知识蒸馏来源的神经机器翻译模型。

需要指出的是，本文提出的方法不仅可以应用于机器翻译任务，而且也具有很好的通用性，可以应用于其他序列到序列的多标签分类任务。未来，除了利用源语言单语数据，我们将探索包括利用目标语言单语数据以及其他多种形式的知识的方法来改进学生神经机器翻译模型。

参考文献

[1] Ilya Sutskever, Oriol Vinyals, Quoc V Le. Sequence to sequence learning with neural networks[C]//Proceedings of the NIPS, 2014: 3104-3112.

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]//Proceedings of the EMNLP, 2014: 1724-1734.

[3] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the ICLR, 2015.

[4] Rico Sennrich, Barry Haddow, Alexandra Birch. Edinburgh neural Machine translation systems for WMT 16[C]//Proceedings of the 1st Conference on Machine Translation, 2016: 371-376.

[5] Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical phrase-based translation[C]//Proceedings of the NAACL, 2003: 48-54.

[6] David Chiang. Hierarchical phrase-based translation [J]. Computational Linguistics, 2007, 33(2): 201-228.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is All You Need[C]//Proceedings of the NIPS, 2017: 6000-6010.

[8] 雷杰，高鑫，宋杰，等. 深度网络模型压缩综述[J]. 软件学报，2018，29(2)：251-266.

[9] Jiaxiang Wu, Cong Leng, Yuhang Wang, et al. Quantized convolutional neural networks for mobile devices [C]//Proceedings of the CVPR, 2016: 4820-4828.

[10] Song Han, Huizi Mao, William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding

[C]//Proceedings of the ICLR, 2016.

[11] Xiyu Yu, Tongliang Liu, Xinchao Wang, et al. On compressing deep models by Low rank and sparse decomposition[C]//Proceedings of the CVPR, 2017: 7370-7379.

[12] Geoffrey Hinton, Oriol Vinyals, Jeffrey Dean. Distilling the knowledge in a neural network[C]//Proceedings of the NIPS Deep Learning and Representation Learning Workshop, 2015.

[13] Guobin Chen, Wongun Choi, Xiang Yu, et al. Learning efficient object detection models with knowledge distillation [C]//Proceedings of the NIPS, 2017: 742-751.

[14] Yoon Kim, Alexander M Rush. Sequence-level knowledge distillation[C]//Proceedings of the EMNLP, 2016: 1317-1327.

[15] Sam Wiseman, Alexander M Rush. Sequence-to-sequence learning as beam-search optimization [C]// Proceedings of the EMNLP, 2016: 1296-1306.

[16] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural machine translation of rare words with subword units [C]//Proceedings of the ACL, 2016: 1715-1725.

[17] Yarin Gal, Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks[C]//Proceedings of the NIPS, 2016: 1019-1027.

[18] Diederik P Kingma, Jimmy Ba. Adam: A method for stochastic optimization [C]//Proceedings of the ICLR, 2015.

[19] Jimmy Lei Ba, Jamie RyanKiros, Geoffrey E Hinton. Layer normalization[J]. arXiv preprint arXiv: 1607.06450, 2016.

[20] Michal Ziemski, Marcin Junczys-Dowmunt, Bruno Pouliquen. The United Nations Parallel Corpus v1.0 [C]//Proceedings of the LREC, 2016.

[21] Cristian Buciluǎ, Rich Caruana, Alexandru Niculescu-Mizil. Model Compression [C]//Proc. KDD, 2006: 535-541.

[22] Jimmy Ba, Rich Caruana. Do deep nets really need to be Deep? [C]//Proceedings of the NIPS, 2014: 2654-2662.

[23] Yun Chen, Yang Liu, Yong Cheng, et al. A teacher-student framework for zero resource neural machine translation [C]//Proceedings of the ACL, 2017: 1925-1935.

[24] Rohan Anil, Gabriel Pereyra, Alexandre Passos, et al. Large scale distributed neural network training through online distillation [C]//Proceedings of the ICLR, 2018.

[25] Jiajun Zhang, Chengqing Zong. Exploiting source-side monolingual data in neural machine translation [C]//Proceedings of the EMNLP, 2016: 1535-1545.

[26] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving neural machine translation models with monolingual data[C]//Proceedings of the ACL, 2016: 86-96.

[27] Yong Cheng, Wei Xu, Zhongjun He, et al. Semi-supervised learning for neural machine translation[C]// Proceedings of the ACL, 2016: 1965-1974.

[28] Tobias Domhan, Felix Hieber. Using target-side monolingual data for neural machine translation through multi-task learning[C]//Proceedings of the EMNLP, 2017: 1500-1505.

李响(1987—)，博士研究生，主要研究领域为自然语言处理。
E-mail: lixiang@ict.ac.cn

陈伟(1984—)，博士，主要研究领域为语音识别。
E-mail: chenweibj8871@sogou-inc.com

刘洋(1979—)，博士，副教授，主要研究领域为自然语言处理。
E-mail: liuyang2011@tsinghua.edu.cn