基于变分信息瓶颈的半监督神经机器翻译

于志强 [1, 2, 3] 余正涛 [1, 3] 黄于欣 [1, 3] 郭军军 [1, 3] 高盛祥 [1, 3]

摘 要 变分方法是机器翻译领域的有效方法, 其性能较依赖于数据量规模. 然而在低资源环境下, 平行语料资源匮乏, 不能满足变分方法对数据量的需求, 因此导致基于变分的模型翻译效果并不理想. 针对该问题, 本文提出基于变分信息瓶颈的半监督神经机器翻译方法, 所提方法的具体思路为: 首先在小规模平行语料的基础上, 通过引入跨层注意力机制充分利用神经网络各层特征信息, 训练得到基础翻译模型; 随后, 利用基础翻译模型, 使用回译方法从单语语料生成含噪声的大规模伪平行语料, 对两种平行语料进行合并形成组合语料, 使其在规模上能够满足变分方法对数据量的需求; 最后, 为了减少组合语料中的噪声, 利用变分信息瓶颈方法在源与目标之间添加中间表征, 通过训练使该表征具有放行重要信息、阻止非重要信息流过的能力, 从而达到去除噪声的效果. 多个数据集上的实验结果表明, 本文所提方法能够显著地提高译文质量, 是一种适用于低资源场景的半监督神经机器翻译方法.

关键词 神经机器翻译, 跨层注意力机制, 回译, 变分信息瓶颈

引用格式 于志强, 余正涛, 黄于欣, 郭军军, 高盛祥. 基于变分信息瓶颈的半监督神经机器翻译. 自动化学报, 2022, **48**(7): 1678−1689

DOI 10.16383/j.aas.c190477

Improving Semi-supervised Neural Machine Translation With Variational Information Bottleneck

YU Zhi-Qiang[1, 2, 3] YU Zheng-Tao[1, 3] HUANG Yu-Xin[1, 3] GUO Jun-Jun[1, 3] GAO Sheng-Xiang[1, 3]

Abstract Variational approach is effective in the field of machine translation, its performance is highly dependent on the scale of the data. However, in low-resource setting, parallel corpus is limited, which cannot meet the demand of variational approach on data, resulting in suboptimal translation effect. To address this problem, we propose a semi-supervised neural machine translation approach based on variational information bottleneck. The central ideas are as follows: 1) cross-layer attention mechanism is introduced to train the basic translation model; 2) the trained basic translation model is used on the basis of small-scale parallel corpus, then get large-scale noisy pseudo-parallel corpus by back-translation with the input of monolingual corpus. Finally, pseudo-parallel and parallel corpora are merged into combinatorial corpora; 3) variational information bottleneck is used to reduce data noise and eliminate information redundancy in the combinatorial corpus. Experiment results on multiple language pairs show that the model we proposed can effectively improve the quality of translation.

Key words Neural machine translation, cross-layer attention mechanism, back-translation, variational information bottleneck

Citation Yu Zhi-Qiang, Yu Zheng-Tao, Huang Yu-Xin, Guo Jun-Jun, Gao Sheng-Xiang. Improving semi-supervised neural machine translation with variational information bottleneck. *Acta Automatica Sinica*, 2022, **48**(7): 1678−1689

自端到端的神经机器翻译 (Neural machine translation) 模型[1−2] 提出以来, 神经机器翻译得到了飞速的发展. 基于注意力机制[2] 的神经机器翻译模型提出之后, 更使得神经机器翻译在很多语言对上的翻译性能超越了传统的统计机器翻译 (Statistical machine translation)[3], 成为自然语言处理领域的热点研究方向[4], 也因此促进了很多神经网络方法在其上的迁移与应用, 变分方法[5−6] 即是其中一种重要方法. 变分方法已证明能够显著提升神经机器翻译的性能[7], 但是由于数据驱动特性, 其性能较

收稿日期 2019-06-24 录用日期 2020-01-17
Manuscript received June 24, 2019; accepted January 17, 2020
国家重点研发计划 (2019QY1800), 国家自然科学基金 (61732005, 61672271, 61761026, 61762056, 61866020), 云南省高新技术产业专项基金 (201606), 云南省自然科学基金 (2018FB104) 资助
Supported by National Key Research and Development Program of China (2019QY1800), National Natural Science Foundation of China (61732005, 61672271, 61761026, 61762056, 61866020), Yunnan High-Tech Industry Development Project (201606), and Natural Science Foundation of Yunnan Province (2018FB104)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 昆明理工大学信息工程与自动化学院 昆明 650500 2. 云南民族大学数学与计算机科学学院 昆明 650500 3. 云南省人工智能重点实验室 昆明 650500
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500 2. School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500 3. Yunnan Key Laboratory of Artificial Intelligence, Kunming 650500

依赖于平行语料的规模与质量, 只有当训练语料规模达到一定数量级时, 变分方法才会体现其优势. 然而, 在低资源语言对上, 不同程度的都面临平行语料缺乏的问题, 因此如何利用相对容易获取的单语语料、实现语料扩充成为应用变分方法的前提. 针对此问题, 本文采用能够同时利用平行语料和单语语料的半监督学习方式展开研究. 半监督神经机器翻译 (Semi-supervised neural machine translation) 主要通过两种方式对单语语料进行利用: 1) 语料扩充−再训练: 利用小规模平行语料训练基础翻译模型, 在此模型基础上利用回译[8] 等语料扩充方法对大规模单语语料进行翻译, 形成伪平行语料再次参与训练; 2) 联合训练: 利用自编码[9−10] 等方法, 以平行语料和单语语料共同作为输入, 进行联合训练. 本文重点关注语料扩充后的变分方法应用, 因此采用语料扩充−再训练方式.

目前被较多采用的语料扩充方法为: 首先利用小规模平行语料训练基础翻译模型, 在此基础上通过回译将大规模单语语料翻译为伪平行语料, 进而组合两种语料进行再次训练. 因此, 基础翻译模型作为任务的起始点, 它的性能直接影响后续任务的执行质量. 传统提升基础翻译模型性能的手段限于使用深层神经网络和在解码端最高层网络应用注意力机制. 然而, 由于深层神经网络在应用于自然语言处理任务中时, 不同层次的神经网络侧重学习的特征不同: 低层网络倾向于学习词法和浅层句法特征, 高层网络则倾向于获取更好的句法结构特征和语义特征[11]. 因此, 很多研究者通过层级注意力机制, 利用神经网络每一层编码器产生的上下文表征指导解码. 层级注意力机制使高层网络的特征信息得以利用的同时, 也挖掘低层网络对输入序列的表征能力. 然而, 上述研究多采用层内融合方式实现层级注意力机制, 其基本方式为将 $k-1$ 层上下文向量融入第 k 层的编码中. 事实上在低资源环境中, 受限的语料规模易导致模型训练不充分, 在此情况下引入层级注意力, 可能会加重网络复杂性, 造成性能下降. 因此, 本文设想通过融入跨层注意力机制, 使低层表征能够跨越层次后对高层表征产生直接影响, 既能弥补因网络复杂性增加带来的性能损失, 又能更好地利用表征信息提升翻译效果. 除此以外, 由于在基础模型的训练过程中缺少双语监督信号, 导致利用其产生的伪平行语料中不可避免的存在大量的数据噪声, 而在增加使用层级注意力机制后, 并不能减少噪声, 相反, 噪声随着更多表征信息的融入呈正比例增长[12−13]. 在随后的再训练过程中, 虽然语料规模能够满足变分方法的需求, 但含有较多噪声的语料作为编码器的输入, 使训练在源头就产生了偏差, 因此对整个再训练过程均造成影响. 针对上述问题, 本文提出了一种融入变分信息瓶颈的神经机器翻译方法. 首先利用小规模平行语料训练得到基础翻译模型, 在其基础上利用回译将大规模单语语料翻译为伪平行语料, 进而合并两种平行语料, 使语料规模达到能够较好地应用变分方法的程度. 在此过程中, 针对基础翻译模型的训练不充分问题, 通过引入跨层注意力机制加强不同层次网络的内部交互, 除了通过注意力机制学习高层网络编码器产生的语义特征之外, 也关注低层网络产生上下文表征的能力和对高层表征的直接影响. 随后, 针对生成的语料中的噪声问题, 使用变分信息瓶颈[12] 方法, 利用其信息控制特性, 在编码端输入 (源语言 x) 与解码端输出 (目标语言 y) 之间的位置引入中间表征, 通过优化中间表征的分布, 使通过瓶颈的有效信息量最大, 从而最大程度放行重要信息、忽略与任务无关的信息, 实现噪声的去除.

本文的创新点包括以下两个方面: 1) 通过融入跨层注意力机制加强基础翻译模型的训练, 在增强的基础翻译模型上利用回译产生伪平行语料、增大数据规模, 使其达到能够有效应用变分方法的程度. 2) 首次将变分信息瓶颈应用于神经机器翻译任务, 在生成的语料的基础上, 利用变分特性提升模型的性能, 同时针对生成语料中的噪声, 利用信息瓶颈的控制特性进行去除. 概括来说, 方法整体实现的是一种语料扩充−信息精炼与利用的过程, 并预期在融合该方法的神经机器翻译中取得翻译效果的提升. 在 IWSLT 和 WMT 等数据集上进行的实验结果表明, 本文提出的方法能显著提高翻译质量.

1 相关工作

1.1 层级注意力机制

注意力机制的有效性得到证明之后, 迅速成为研究者们关注的热点. 很多研究者在神经网络的不同层次上应用注意力机制构建层级注意力模型, 在此基础上展开训练任务. Yang 等[14] 将网络划分为两个注意力层次, 第一个层次为 “词注意”, 另一个层次为 “句注意”, 每部分通过双向循环神经网络 (Recurrent neural network) 结合注意力机制实现文本分类. Pappas 等[15] 提出了一种用于学习文档结构的多语言分层注意力网络, 通过跨语言的共享编码器和注意力机制, 使用多任务学习和对齐的语义空间作为文本分类任务的输入, 显著提升分类效果. Zhang 等[16] 提出一种层次结构摘要方法, 使用分层结构的自我关注机制来创建句子和文档嵌入, 通过层次注意机制提供额外的信息源来获取更佳的特征表示, 从而更好地指导摘要的生成. Miculi-

cich 等[17] 提出了一个分层关注模型, 将其作为另一个抽象层次集成在传统端到端的神经机器翻译结构中, 以结构化和动态的方式捕获上下文, 显著提升了结果的 BLEU (Bilingual evaluation understudy) 值. Zhang 等[18] 提出了一种深度关注模型, 模型基于低层网络上的注意力信息, 自动确定从相应的编码器层传递信息的阈值, 从而使词的分布式表示适合于高层注意力, 在多个数据集上验证了模型的有效性. 研究者们通过融入层级注意力机制到模型训练中, 在模型之上直接执行文本分类、摘要和翻译等任务, 与上述研究工作不同的是, 本文更关注于跨层次的注意力机制, 并期待将融入跨层注意力机制的基础翻译模型用于进一步任务.

1.2 单语语料扩充

如何在低资源场景下进行单语语料的扩充和利用一直是研究者们关注的热点问题之一. 早在 2007 年, Ueffing 等[19] 就提出了基于统计机器翻译的语料扩充方法: 利用直推学习来充分利用单语语料库. 他们使用训练好的翻译模型来翻译虚拟的源文本, 将其与译文配对, 形成一个伪平行语料库. 在此基础上, Bertoldi 等[20] 通过改进的网络结构进行训练, 整个过程循环迭代直至收敛, 取得了性能上的进一步提升. Klementiev 等[21] 提出了一种单语语料库短语翻译概率估计方法, 在一定程度上缓解了生成的伪平行语料中的重复问题. 与前文不同, Zhang 等[22] 使用检索技术直接从单语语料库中提取平行短语. 另一个重要的研究方向是将基于单语语料库的翻译视为一个解密问题, 将译文的生成过程等同于密文到明文的转换[23−24].

以上的单语语料扩充方法主要应用于统计机器翻译中. 随着深度学习的兴起, 神经机器翻译成为翻译任务的主流方法, 探索在低资源神经机器翻译场景下的语料扩充方法成为研究热点. Sennrich 等[8] 在神经机器翻译框架基础上提出了语料扩充方法. 他们利用具有网络结构普适性的两种方法来使用单语语料. 第 1 种方法是将单语句子与虚拟输入配对, 然后在固定编码器和注意力模型参数的情况下利用这些伪平行句对进行训练. 在第 2 种方法中, 他们首先在平行语料库上训练初步的神经机器翻译模型, 然后使用该模型翻译单语语料, 最后结合单语语料及其翻译构成伪平行语料, 第 2 种方法也称为回译. 回译可在不依赖于神经网络结构的情况下实现平行语料的构建, 因此广泛应用于半监督和无监督神经机器翻译中. Cheng 等[25] 提出一种半监督神经机器翻译模型, 通过将回译与自编码进行结合重构源与目标语言的伪平行语料, 取得了翻译性能上的提升. Skorokhodov 等[26] 提出了一种将知识从单独训练的语言模型转移到神经机器翻译系统的方法, 讨论了在缺乏平行语料和计算资源的情况下, 利用回译等方法提高翻译质量的几种技术. Artetxe 等[27] 利用共享编码器, 在两个解码器上分别应用回译与去噪进行联合训练, 实现了只依赖单语语料的非监督神经机器翻译. Lample 等[28]提出了两个模型变体: 一个神经网络模型和一个基于短语的模型. 利用回译、语言模型去噪以及迭代反向翻译自动生成平行语料. Burlot 等[29] 对回译进行了系统研究, 并引入新的数据模拟模型实现语料扩充. 与上述研究工作不同的是, 本文同时关注于伪平行语料生成所依赖的基础翻译模型的训练. 在训练过程中, 不仅利用注意力机制关注高层网络中对句法结构和语义信息的利用, 同时也关注低层网络信息对高层网络信息的直接影响.

1.3 变分信息瓶颈

为了实现信息的压缩和去噪, Tishby 等[30] 提出基于互信息的信息瓶颈 (Information bottleneck) 方法. 深度神经网络得到广泛应用后, Alemi 等[12] 在传统信息瓶颈的基础上进行改进, 提出了适用于神经网络的变分信息瓶颈 (Variational information bottleneck), 变分信息瓶颈利用深度神经网络来建模和训练, 通过在源和目标之间添加中间表征来进行信息过滤.

在神经机器翻译中, 尚未发现利用变分信息瓶颈进行噪声去除的相关研究工作, 但是一些基于变分的方法近期已经在神经机器翻译中得到应用, 有效提高了翻译性能. Zhang 等[7] 提出一个变分模型, 通过引入一个连续的潜在变量来显式地对源语句的底层语义建模并指导目标翻译的生成, 能够有效的提高翻译质量. Eikema 等[31] 提出一个双语句对的深层生成模型, 该模型从共享的潜在空间中共同生成源句和目标句, 通过变分推理和参数梯度化来完成训练, 在域内、混合域等机器翻译场景中证明了模型的有效性. Su 等[32] 基于变分递归神经网络, 提出了一种变分递归神经机器翻译模型, 利用变分自编码器将随机变量添加到解码器的隐藏状态中, 能够在不同的时间步长上进一步捕获依赖关系.

2 模型

本节首先介绍传统基于注意力机制的基础翻译模型, 接着介绍了融入跨层注意力机制的基础翻译模型. 区别于传统的基础翻译模型, 本文通过融入跨层注意力机制, 除关注高层编码器产生的上下文

表征向量之外, 也关注低层编码器产生的上下文表征向量对高层编码的直接影响. 最后介绍了变分信息瓶颈模型, 展示了利用该模型对回译方法生成的伪平行语料中的噪声进行去除的过程.

2.1　传统注意力机制模型

传统方法中, 最初通过在解码端最高层网络引入注意力机制进行基础翻译模型的训练. 如图 1 所示的 2 层编解码器结构中, 它通过在每个时间步长生成一个目标单词 y_t 来进行翻译. 给定编码端输入序列 $x=(x_1, x_2, \cdots, x_n)$和已生成的翻译序列 $y=(y_1, y_2, \cdots, y_{t-1})$, 解码端产生下一个词 y_t 的概率为

$$P(y_t|y_{<t}, x)=\text{softmax}(g(y_{t-1}, s_t, c_t)) \tag{1}$$

其中, g 是非线性函数, s_t 为在时间步 t 时刻的解码端隐状态向量, 由下式计算得到

$$s_t=f(s_{t-1}, y_{t-1}, c_t) \tag{2}$$

其中, f 是激活函数, c_t 是 t 时刻的上下文向量, 其计算式为

$$c_t=\sum_{j=1}^{T_x}\alpha_{t,j}h_j \tag{3}$$

其中, $h_j=[\overrightarrow{h}_j; \overleftarrow{h}_j]$ 是输入序列 x_j 的向量表征, 由前向和后向编码向量拼接得到. 权重 $\alpha_{t,j}$ 的定义为

$$\alpha_{t,j}=\frac{\exp(e_{t,j})}{\sum_{k=1}^{T_x}\exp(e_{t,k})} \tag{4}$$

其中, $e_{t,j}$ 是对 s_{t-1} 和 h_j 相似性的度量, 其计算式为

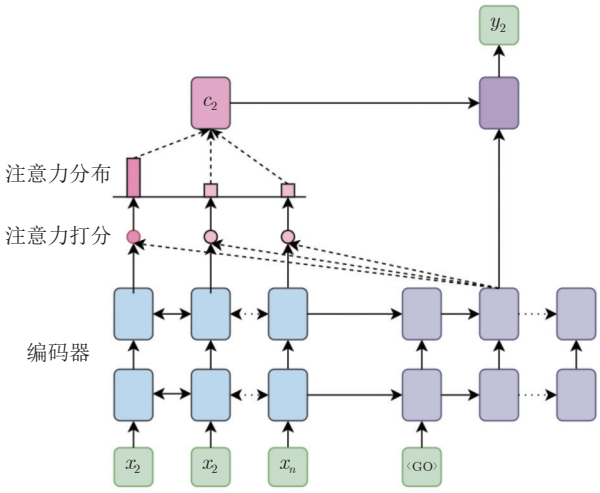


图 1　传统作用于最高层网络的注意力机制融入

Fig. 1　Model with traditional attention mechanism based on top-layer merge

$$e_{t,j}=a(s_{t-1}, h_j) \tag{5}$$

通过在最高层网络引入注意力机制来改善语义表征、辅助基础翻译模型的训练, 能够有效地提升翻译性能, 但仅利用最高层信息的方式使得其他层次的词法和浅层句法等特征信息被忽略, 进而影响生成的伪平行语料质量. 针对此问题, 能够利用每层网络上下文表征的层级注意力机制得到关注, 成为众多翻译系统采用的基础方法. 这些系统往往采用层内融合方式的层级注意力机制, 如图 2 所示的编解码器结构中, 第 k 层的输入融合了 $k-1$ 层的上下文向量和隐状态向量. 具体计算式为

$$d_t^{k-1}=\tanh(W_d[s_t^{k-1}; c_t^{k-1}]+b_d) \tag{6}$$

$$s_t^k=f(s_{t-1}^k, d_t^{k-1}) \tag{7}$$

$$p_t=\tanh(W_p([s_t^r; c_t^r])+b_p) \tag{8}$$

其中, f 为激活函数, r 为神经网络层数.

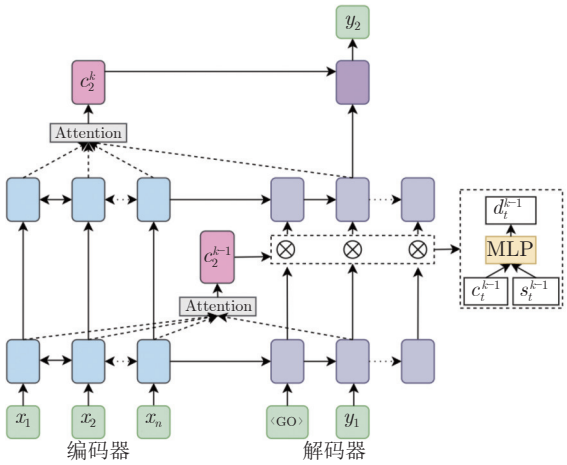


图 2　层内融合方式的层级注意力机制融入

Fig. 2　Model with hierarchical attention mechanism based on inner-layer merge

2.2　跨层注意力机制模型

层内融合方式加强了低层表征利用, 但难以使低层表征跨越层次对高层表征产生直接影响. 因此, 本文设想利用跨层融合, 在利用低层表征的同时促进低层表征对高层表征的直接影响. 通过融入跨层注意力机制, 使各层特征信息得到更加充分的利用. 如图 3 所示, 模型通过注意力机制计算每一层的上下文向量 c_t^k, 在最高层 r 对它们进行拼接, 得到跨层融合的上下文向量 c_t

$$c_t=[c_t^1; c_t^2; \cdots; c_t^r] \tag{9}$$

同样, 通过跨层拼接操作得到 s_t, 随后通过非

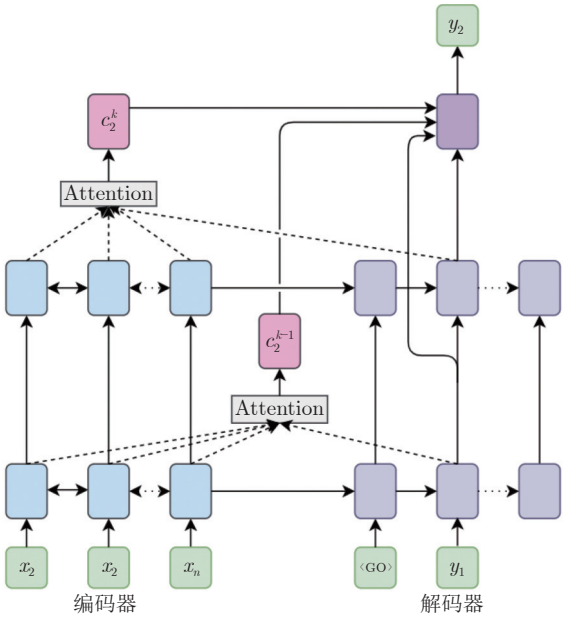


图 3 跨层融合方式的层级注意力机制融入

Fig. 3 Model with hierarchical attention mechanism based on cross-layer merge

线性变换得到 p_t, p_t 用于输入到softmax 函数中计算词表中的概率分布

$$s_t = [s_t^1; s_t^2; \cdots; s_t^r] \tag{10}$$

$$p_t = \tanh(W_p([s_t; c_t]) + b_p) \tag{11}$$

2.3 变分信息瓶颈模型

在基础翻译模型的训练中, 通过融入不同层次的上下文向量来改善语义表征, 但也因此带来更多的噪声信息. 针对此问题, 本文通过在编解码结构中引入适用于神经网络的变分信息瓶颈方法来进行解决. 需要注意的是, 编解码结构中, 编码端的输入通过编码端隐状态隐式传递到解码端. 变分信息瓶颈要求在编码端输入与解码端最终输出之间的位置引入中间表征, 因此为了便于实现, 将变分信息瓶颈应用于解码端获取最终输出之前, 以纳入损失计算的方式进行模型训练, 其直接输入为解码端的隐状态, 以此种方式实现对编码端输入中噪声的过滤. 具体流程为: 在给定的 X 到 Y 的转换任务中, 引入 Z 作为源输入 X 的中间表征, 构造从 $X \to Z \to Y$ 的信息瓶颈 $R_{IB}(\theta)$, 利用 Z 实现对 X 中信息的筛选和过滤. 计算过程为

$$I(Z, Y; \theta) = \int p(z, y|\theta) \log \frac{p(z, y|\theta)}{p(z|\theta)p(y|\theta)} \mathrm{d}x\mathrm{d}y \tag{12}$$

$$R_{IB}(\theta) = I(Z, Y; \theta) - \beta I(Z, X; \theta) \tag{13}$$

其中, $I(Z, Y; \theta)$ 表示 Y 和 Z 之间的互信息量. 变分信息瓶颈的目标是以互信息作为信息量的度量, 通过学习编码 Z 的分布, 使 $X \to Y$ 的信息量最小, 强迫模型让最重要的信息流过信息瓶颈而忽略与任务无关的信息, 从而实现噪声的去除.

给定输入平行语料 $D = \{\langle x^{(n)}, y^{(n)}\rangle\}_{n=1}^{N}$, 神经机器翻译的标准训练目标是极大化训练数据的似然概率

$$L(\theta) = \sum_{n=1}^{N} \log P\left(y^{(n)}|x^{(n)}; \theta\right) \tag{14}$$

其中, $P(y|x; \theta)$ 是 $x \to y$ 的翻译模型, θ 为模型的参数集合. 训练过程中, 寻求极大化似然概率等价于寻求损失的最小化

$$loss = -\log P(y|x; \theta) \tag{15}$$

本文引入信息瓶颈 $z = f(x, y_{<t})$ 作为编码的中间表征, 构造从中间表征 z 到输出序列 y 的损失, 作为训练的交叉熵损失, 计算式为

$$P(y|z;) = \sum_{t=1}^{N} \log P(y_t|z; \theta) \tag{16}$$

同时加入约束, 目标为 $P(z|x; \theta)$ 的分布与标准正态分布 $Q(z)$ 的 KL 散度 (Kullback-Leibler divergence) 最小化, 在引入变分信息瓶颈之后, 训练过程的损失函数为

$$loss_{\mathrm{VIB}} = -\log P(y, z|x; \theta_1) + \lambda KL(P(z|x; \theta_2) \parallel Q(z)) \tag{17}$$

其中, λ 为超参数, 实验结果表明, λ 设置为 10^{-3} 时取得最优结果.

图 4 显示了引入了变分信息瓶颈后的模型结构, 同样地, 为了利用不同层次的上下文表征信息, 在变分信息瓶颈模型中也引入了跨层注意力机制. 模型的输入为平行语料和伪平行语料的组合. 以给定小规模平行语料 $D_{a,b} = \{\langle a^{(m)}, b^{(m)}\rangle\}_{m=1}^{M}$ 和单语语料 $D_x = \{\langle x^{(n)}\rangle\}_{n=1}^{N}$ 为例, 表 1 展示了由原始小规模平行语料 $D_{a,b}$ 和由单语语料 D_x 生成的伪平行语料 $D_{x,y}$ 进行组合, 形成最终语料 $D_{b+y,a+x} = \{\langle b^{(m)} + y^{(n)}, a^{(m)} + x^{(n)}\rangle\}_{m,n=1}^{M+N}$ 的过程. 需要注意的是, 变分信息瓶颈是通过引入中间表征来实现去除源输入中的噪声信息, 对于单语语料 $D_x = \{\langle x^{(n)}\rangle\}_{n=1}^{N}$ 而言, 噪声信息存在于通过回译生成的对应伪语料 $D_y = \{\langle y^{(n)}\rangle\}_{n=1}^{N}$ 中. 因此在模型训练时, 需调换翻译方向, 将包含噪声信息的 $D_b + D_y$ 作为源语言语料进行输入, 对其进行噪声去除. 而目标语言为不含噪声的 $D_a + D_x$, 利于损失的计算.

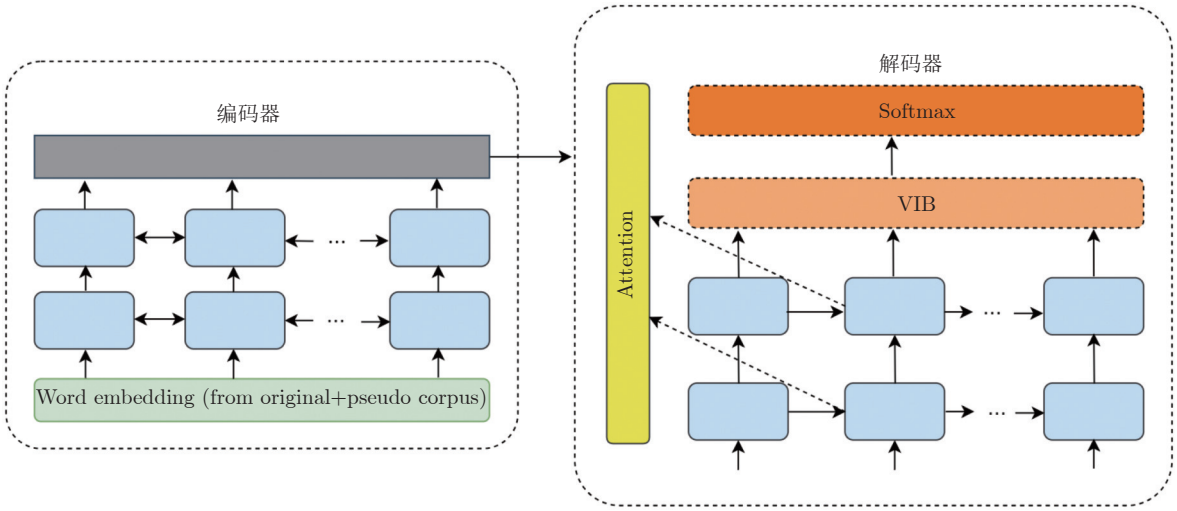


图 4　融入变分信息瓶颈后的神经机器翻译模型

Fig. 4　NMT model after integrating variational information bottleneck

表 1　语料组合结构示例

Table 1　Examples of the combined corpus structure

语料类别	源语言语料	目标语言语料
原始语料	D_a	D_b
单语语料	D_x	—
伪平行语料	D_x	D_y
组合语料	$D_b + D_y$	$D_a + D_x$

3　实验设置

3.1　数据集

本文选择机器翻译领域的通用数据集作为平行语料来源, 表 2 显示了平行语料的构成情况. 为观察本文方法在不同规模数据集上的作用, 采用不同规模的数据集进行对比实验. 小规模训练语料中, 英–越、英–中和英–德平行语料均来自 IWSLT15 数据集, 本文选择 tst2012 作为验证集进行参数优化和模型选择, 选择 tst2013 作为测试集进行测试验证. 大规模训练来自 WMT14 数据集, 验证集和测试集分别采用 newstest2012 和 newstest2013.

表 2　平行语料的构成

Table 2　The composition of parallel corpus

语料类型	数据集	语言对	训练集	验证集	测试集
小规模平行语料	IWSLT15	en ↔ vi	133 K	1553	1268
	IWSLT15	en ↔ zh	209 K	887	1261
	IWSLT15	en ↔ de	172 K	887	1565
大规模平行语料	WMT14	en ↔ de	4.5 M	3003	3000

注: en: 英语, vi: 越南语, zh: 中文, de: 德语.

表 3 显示了单语语料的构成情况, 英–越和英–中翻译中, 英文和中文使用的单语语料来源于 GIGAWORD 数据集, 越南语方面为互联网爬取和人工校验结合处理后得到的 1 M 高质量语料. IWSLT 和 WMT 上的英–德翻译任务中, 使用的单语语料来源于 WMT14 数据集的单语部分, 具体由 Europarl v7、News Commentary 和 News Crawl 2011 组合而成. 本文对语料进行标准化预处理, 包括词切分、过长句对过滤, 其中, 对英语、德语还进行了去停用词操作. 本文选择 BPE 作为基准系统, 源端和目标端词汇表大小均设置为 30 000.

表 3　实验使用的单语语料的构成, 其中越南语使用本文构建的单语语料

Table 3　The composition of monolingual corpus, in which Vietnamese was collected by ourselves

	翻译任务	语言	数据集	句数 (M)
单语语料	en ↔ vi	en	GIGAWORD	22.3
		vi	None	1
	en ↔ zh	en	GIGAWORD	22.3
		zh	GIGAWORD	18.7
	en ↔ de (IWSLT15)	en	WMT14	18
		de	WMT14	17.3
	en ↔ de (WMT14)	en	WMT14	18
		de	WMT14	17.3

3.2　参数设置

本文选择以下模型作为基准系统:

1) RNNSearch 模型: 编码器和解码器分别采

用 6 层双向长短期记忆网络 (Bi-directional long short-term memory, Bi-LSTM) 和长短期记忆网络 (Long short-term memory, LSTM) 构建. 隐层神经元个数设置为 1 000, 词嵌入维度设置为 620. 使用 Adam 算法[33] 进行模型参数优化, dropout 率设定为 0.2, 批次大小设定为 128. 使用集束宽度为 4 的集束搜索 (Beam search) 算法进行解码.

2) Transformer 模型: 编码器和解码器分别采用默认的 6 层神经网络, 头数设置为 8, 隐状态和词嵌入维度设置为 512. 使用 Adam 算法进行模型参数优化, dropout 率设定为 0.1, 批次大小设置为 4096. 测试阶段使用集束搜索算法进行解码, 集束宽度为 4.

利用 IWSLT15 数据集进行的小规模平行语料实验中, 本文参考了 Sennrich 等[34] 关于低资源环境下优化神经机器翻译效果的设置, 包括层正则化和激进 dropout.

3.3 评价指标

本文选择大小写不敏感的 BLEU 值[35] 作为评价指标, 评价脚本采用大小写不敏感的 multi-bleu.perl. 为了从更多角度评价译文质量, 本文另外采用 RIBES 进行辅助评测. RIBES (Rank-based intuitive bilingual evaluation score) 是另一种评测机器翻译性能的方法[36], 与 BLEU 评测不同的是, RIBES 评测方法侧重于关注译文的词序是否正确.

4 实验结果分析

本节首先通过机器翻译评价指标对提出的模型进行量化评价, 接着通过可视化的角度对模型效果进行了分析.

4.1 BLEU 值评测

本文提出的方法和基准系统在不同翻译方向上的 BLEU 值如表 4 所示, 需要注意的是, 为了应用变分信息瓶颈、实现对源端噪声信息进行去除, 最终翻译方向与基础翻译模型方向相反 (具体原因见第 2.3 节中对表 1 的描述). 表 4 中 RNNSearch 和 Transformer 为分别在基线系统上, 利用基础模型进行单语语料回译, 接着将获得的组合语料再次进行训练后得到的 BLEU 值. 表 4 同时展示了消融不同模块后的 BLEU 值变化, 其中 CA、VIB 分别表示跨层注意力、变分信息瓶颈模块.

通过实验结果可以观察到, 本文提出的融入跨层注意力和变分信息瓶颈方法在所有翻译方向上均取得了性能提升. 以在 IWSLT15 数据集上的德→英翻译为例, 相较 Transformer 基准系统, 融入两种方法后提升了 0.69 个 BLEU 值. 同时根据德英翻译任务结果可以观察到, BLEU 值的提升幅度随着语料规模的上升而减小. 出现该结果的一个可能原因是在低资源环境下, 跨层注意力的使用能够挖掘更多的表征信息、使低层表征对高层表征的影响更为直接. 而在资源丰富的环境下, 平行语料规模提升所引入的信息与跨层注意力所挖掘信息在一定程度上有所重合. 另一个可能原因是相对于资源丰富环境, 低资源环境产生的伪平行语料占组合语料的比例更大, 变分信息瓶颈进行了更多的噪声去除操作.

表 4 BLEU 值评测结果 (%)
Table 4 Evaluation results of BLEU (%)

模型	BLEU							
	en→vi	vi→en	en→zh	zh→en	en→de (IWSLT15)	de→en (IWSLT15)	en→de (WMT14)	de→en (WMT14)
RNNSearch	26.55	24.47	21.18	19.15	25.03	28.51	26.62	29.20
RNNSearch+CA	27.04	24.95	21.64	19.59	25.39	28.94	27.06	29.58
RNNSearch+VIB	27.35	25.12	21.94	19.84	25.77	29.31	27.27	29.89
RNNSearch+CA+VIB	**27.83***	**25.61***	**22.39**	**20.27**	**26.14***	**29.66***	**27.61***	**30.22***
△	**+1.28**	**+1.14**	**+1.21**	**+1.12**	**+1.11**	**+1.15**	**+0.99**	**+1.02**
Transformer	29.20	26.73	23.69	21.61	27.48	30.66	28.74	31.29
Transformer+CA	29.53	27.00	23.95	21.82	27.74	30.98	28.93	31.51
Transformer+VIB	29.96	27.38	24.30	22.13	28.04	31.24	29.16	31.75
Transformer+CA+VIB	**30.17***	**27.56***	**24.43**	**22.32**	**28.11***	**31.35***	**29.25***	**31.89***
△	**+0.97**	**+0.83**	**+0.74**	**+0.71**	**+0.63**	**+0.69**	**+0.51**	**+0.60**

注: △ 表示融入CA+VIB后相较基准系统的BLEU值提升, * 表示利用bootstrap resampling[37] 进行了显著性检验 ($p < 0.05$)

上述实验证明了本文所提方法可以融入不同框架使用, 同时适用于资源丰富环境和低资源环境, 尤其在平行语料匮乏的低资源环境下, 能够通过充分利用神经网络各层信息来加大信息量, 同时通过去噪改善信息的质量, 在某种程度上与增加高质量平行语料具有同类效果.

为了观察单独或共同融入跨层注意力和变分信息瓶颈后在不同翻译方向上 BLEU 值的提升效果, 本文采用消融方式单独或组合使用两种方法并在表 4 中报告了实验结果. 以 IWSLT15 数据集上的英→越翻译为例, 相对于 Transformer 基准系统, 单独使用跨层注意力在测试集上获得了 0.33 个 BLEU 值的提升, 而单独使用变分信息瓶颈提高了 0.76 个 BLEU 值, 结合使用两种方法后则提高了 0.97 个 BLEU 值.

消融实验结果表明, 本文所提两种方法既可以独立地应用于翻译框架中, 也可以结合使用. 单独融入时均能带来一定的翻译性能上的提升, 而结合使用两种方法后, 则获得了进一步的提升. 因此我们认为, 结合使用跨层注意力和变分信息瓶颈能够在有效加大信息量的同时提升翻译质量.

此外, 将本文所提方法与 Zhang 等[38] 提出的基于回译的半监督联合训练方法进行了对比实验. 该方法以联合学习方式, 在每步迭代中首先利用回译得到伪平行语料, 将其投入训练后通过极大似然估计交替优化源到目标和目标到源的翻译模型, 实现了翻译性能的较大提升. 为保持标准一致, 实验语料和训练参数均沿用该文设置: 平行语料为 WMT14 英德训练集 (4.5 M), 单语语料截取自 News Crawl 2012 (8 M); 选取 RNNSearch 为基础模型, 隐层神经元个数设置为 1 024, 词嵌入维度设置为 256, 批次大小设定为 128. 本文方法中, 同时使用了跨层注意力和变分信息瓶颈机制.

通过表 5 的实验结果可以观察到, 相较于对比方法, 本文所提方法在英–德和德–英方向上取得了 1.13 和 0.67 个 BLEU 值的提升.

表 5　与其他半监督方法的比较 (en-de)

Table 5　Comparison between our work and different semi-supervised NMT approach (en-de)

模型	翻译方向	基础翻译模型	单语语料	BLEU
Zhang et al. (2018)	en→de	de→en	de	23.60
	de→en	en→de	en	27.98
this work	en→de	de→en	de	24.73
	de→en	en→de	en	28.65

4.2　RIBES 值评测

本文利用 RIBES 方法对 IWSLT 数据集中 2 个语言对评测结果如表 6 所示, 其中基准模型为实现层内注意力机制的 Transformer 模型. 从表中可以观察到, 相较基准系统, 融入跨层注意力机制后在所有翻译任务上均取得了 RIBES 值的提升, 其中英→越翻译任务上, 提升了最高的 0.69 个 RIBES 值, 在此基础上使用变分信息瓶颈模型, 则获得了 1.45 个 RIBES 值的提升. 因此实验结果表明, 相较基准系统, 融入跨层注意力机制可完善译文的句子结构信息, 起到了词序优化的作用. 在此基础上结合使用变分信息瓶颈方法, 生成的译文则具有更佳的词序.

表 6　RIBES 值评测结果 (%)

Table 6　Evaluation results of RIBES (%)

翻译方向	基础翻译模型	单语语料	基准模型	跨层注意力	跨层注意力+变分信息瓶颈
en→vi	vi→en	vi	74.38	75.07	75.83
vi→en	en→vi	en	74.29	74.70	75.64
en→zh	zh→en	zh	72.87	73.33	73.83
zh→en	en→zh	en	71.81	72.25	72.55
en→de (IWSLT15)	de→en	de	79.81	80.14	80.96
de→en (IWSLT15)	en→de	en	78.48	78.88	79.61
en→de (WMT14)	de→en	de	80.15	80.40	81.29
de→en (WMT14)	en→de	en	79.33	79.52	80.07

4.3　句子结构缺失问题

为了更直观地验证融入跨层注意力机制是否更能促进高层网络的句法结构信息完善, 同时验证融入变分信息瓶颈是否具有噪声信息去除的作用, 本文在中–英翻译结果中随机选取了 300 句译文, 并对译文质量进行分析.

表 7 展示了一个翻译实例 (TA、CA、CA+VIB 分别表示传统注意力、跨层注意力、跨层注意力加变分信息瓶颈方法): 给定中文源句 “火车被发现已经开走了”, 作为基准系统的传统注意力模型产生的译文为 “Found that the train had gone”, 句中缺少形式主语造成句子结构并不完整. 而融入跨层注意力机制后的译文为 “It was found that the the train had left away”, 缺失的形式主语被补全, 句子结构得到了完善. 通过对全部 300 句译文进行分析发现, 融入跨层注意力机制后产生的译文, 其结

构完整性普遍增强. 因此, 实验证明, 使用跨层注意力机制, 在学习到低层网络所产生的词法和浅层句法特征同时, 对学习高层网络所产生的句法结构特征也具有促进作用, 有助于提升句子结构的完整性.

表 7 中–英翻译实例

Table 7 Chinese-English translation examples

源句	参考译文	真实译文
火车被发现已经开走了	It was found that the train had already left	[TA] Found that the the train had gone [CA] It was found that the the train had left away [CA+VIB] It was found that the train had left

4.4 过度翻译问题

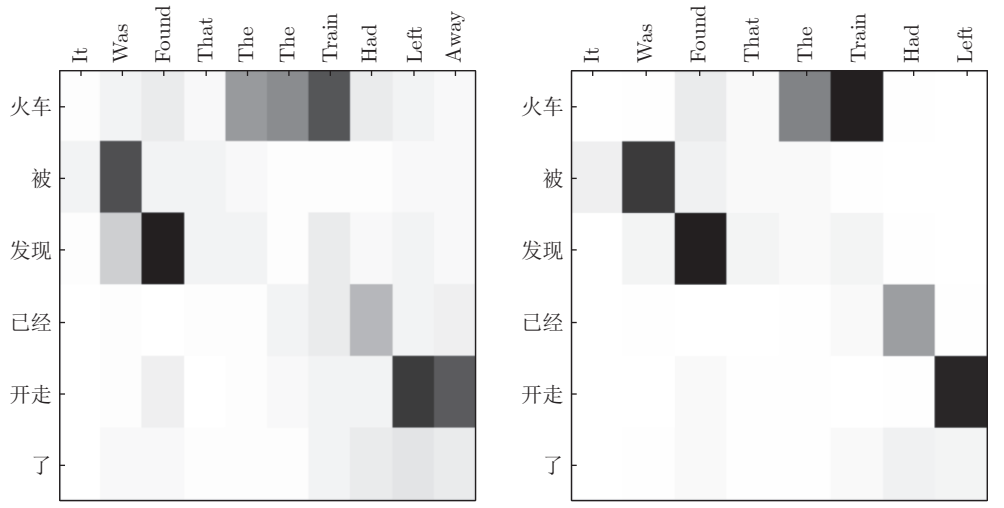
神经机器翻译中, 源端语料中的噪声信息会导致过度翻译问题. 在本文讨论的低资源场景下, 利用回译构建伪平行语料实现语料扩充, 但由于在过程中缺乏足够的监督信号, 导致语料规模扩充的同时产生了噪声信息, 进而引发过度翻译问题. 通过第 4.3 节的实验发现, 融入跨层注意力机制可以缓解句子结构缺失问题, 但未能消除句子中存在的噪声信息. 通过表 7 中的示例可以观察到: 利用传统方法得到的译文中, “the” 被重复翻译了一次, 而在使用跨层注意力后, 重复的 “the” 并没有被消除; 同时虽然 “gone” 被翻译为更合理的 “left”, 但也产生了多余的译文 “away”. 因此, 跨层注意力机制虽然能够通过使用更多的特征信息来提升句子的完整性, 但并未解决的噪声信息问题. 针对该问题, 本文在融入跨层注意力机制后应用变分信息瓶颈模型, 并对结果进行可视化处理. 如图 5 所示, 可以观察到图 5(a) 中冗余的 “the” 和 “away” 在图 5(b) 中被去除. 除此以外, 在图 5(a) 中, 由于噪声信息 “the” 和 “away” 的出现, 使得 “train” 和 “left” 等译文的关注度相对分散; 而通过图 5(b) 可以观察到, 在消除了噪声信息后, 译文的关注更加集中.

4.5 译文长度

本文将测试集中的源语言句子按长度分为 8 组, 然后评测翻译产生的相应译文的长度. 英–越翻译任务的译文长度评测结果如图 6 所示, 从图 6(a) 中观察到, 英→越翻译方向上, 只使用跨层注意力机制 (CA) 后, 基础翻译模型所产生的译文长度优于作为基准系统的传统模型 (TA). 而在此基础上应用变分信息瓶颈后 (CA+VIB), 在区间 [10, 60] 内的译文长度高于基准系统, 在其余长度区间内低于基准系统. 实验表明, 跨层注意力机制通过引入多层特征, 能有效地提升译文句长. 而在应用变分信息瓶颈模型进行噪声信息过滤后, 译文长度下降, 但在译文的常规分布区间 [10, 60] 内仍优于基准系统.

4.6 参数对模型的影响

变分信息瓶颈中的超参数 λ 对损失的求解产生直接的影响, 因此本文在表 4 所示的不同的翻译方向展开实验, 分析了 λ 取不同值时的翻译效果, 图 7



(a) 融入跨层注意力后仍存在过度翻译问题

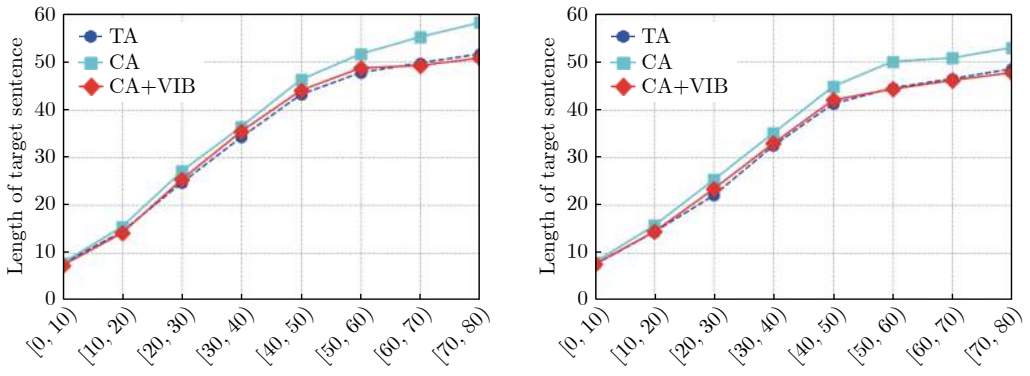
(a) Excessive translation problem still exists after integrating cross-layer attention

(b) 变分信息瓶颈缓解了过度翻译问题

(b) Variational information bottleneck alleviates the excessive translation problem

图 5 翻译效果可视化

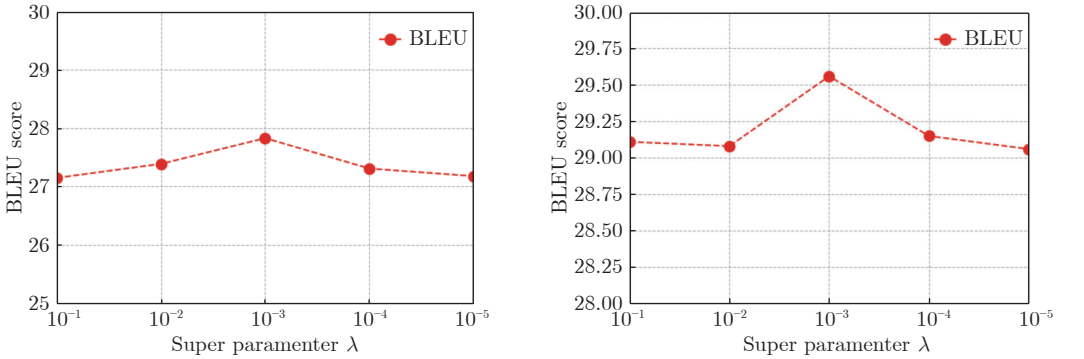
Fig. 5 Example of translation effects



(a) en→vi　　(b) vi→en

图 6　英–越翻译任务的译文长度评测

Fig. 6　Translation length evaluation of English-Vietnamese translation task



(a) en→vi 方向 λ 对 BLEU 值的影响 (RNNSearch)
(a) The influence of λ on BLEU value for en→vi translation task (RNNSearch)

(b) en→vi 方向 λ 对 BLEU 值的影响 (Transformer)
(b) The influence of λ on BLEU value for en→vi translation task (Transformer)

图 7　λ 参数对模型的影响

Fig. 7　Influence of λ parameter on the model

展示了在英→越翻译上 λ 对 BLEU 值的影响. 综合考虑所有翻译任务, λ 取 0.001 时的翻译质量最好.

5　结束语

传统的回译模型侧重于关注产生的伪平行语料规模, 在生成基础翻译模型时, 缺乏对神经网络跨层次信息的重视. 在使用深层神经网络进行初步模型训练时, 仅局限于利用最高层或各层内部的语义信息作为上下文表征, 忽略了低层网络对高层网络表征的直接促进作用, 因此对句法结构等信息表征不足, 造成伪平行语料生成过程中的信息缺失. 针对此问题, 本文首先通过引入跨层注意力机制加强对各层网络信息的利用, 随后基于此基础训练模型进行语料扩充, 使语料在规模上能够满足变分方法的应用需求. 然而, 跨层注意力机制在加强特征信息利用、改善基础翻译模型的同时, 会进一步引入噪声信息, 针对此问题, 本文通过引入变分信息瓶颈来进行噪声的消除. 在多个翻译数据集上的实验结果表明, 相较基准系统, 本文提出的方法在有效提高译文质量的同时保持了译文句长, 并在一定程度上解决了传统神经机器翻译中出现的过度翻译问题.

References

1　Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 3104–3112

2　Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA, 2015. 1–15

3　Jiang Hong-Fei, Li Sheng, Zhang Min, Zhao Tie-Jun, Yang Mu-Yun. Synchronous tree sequence substitution grammar for statistical machine translation. *Acta Automatica Sinica*, 2009, **35**(10): 1317–1326
(蒋宏飞, 李生, 张民, 赵铁军, 杨沐昀. 基于同步树序列替换文法的

统计机器翻译模型. 自动化学报, 2009, **35**(10): 1317−1326)

4 Li Ya-Chao, Xiong De-Yi, Zhang Min. A survey of neural machine translation. *Chinese Journal of Computers*, 2018, **41**(12): 2734−2755
(李亚超, 熊德意, 张民. 神经机器翻译综述. 计算机学报, 2018, **41**(12): 2734−2755)

5 Kingma D P, Rezende D J, Mohamed S, Welling M. Semi-supervised learning with deep generative models. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 3581−3589

6 Kingma D P, Welling M. Auto-encoding variational bayes. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, Canada, 2014.

7 Zhang B, Xiong D Y, Su J S, Duan H, Zhang M. Variational neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). Austin, USA: Association for Computational Linguistics, 2016. 521−530

8 Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016. 86−96

9 Socher R, Pennington J, Huang E H, Ng A Y, Manning C D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, UK: Association for Computational Linguistics, 2011. 151−161

10 Ammar W, Dyer C, Smith N A. Conditional random field autoencoders for unsupervised structured prediction. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 3311−3319

11 Belinkov Y, Durrani N, Dalvi F, Sajjad H, Glass J. What do neural machine translation models learn about morphology? In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 861−872

12 Alemi A A, Fischer I, Dillon J V, Murphy K. Deep variational information bottleneck. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: OpenReview.net, 2017.

13 Nguyen T T, Choi J. Layer-wise learning of stochastic neural networks with information bottleneck. arXiv: 1712.01272, 2017.

14 Yang Z C, Yang D Y, Dyer C, He X D, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics, 2016. 1480−1489

15 Pappas N, Popescu-Belis A. Multilingual hierarchical attention networks for document classification. In: Proceedings of the 8th International Joint Conference on Natural Language Processing. Taipei, China: Asian Federation of Natural Language Processing, 2017. 1015−1025

16 Zhang Y, Wang Y H, Liao J Z, Xiao W D. A hierarchical attention Seq2seq model with CopyNet for text summarization. In: Proceedings of the 2018 International Conference on Robots and Intelligent System (ICRIS). Changsha, China: IEEE, 2018. 316−320

17 Miculicich L, Ram D, Pappas N, Henderson J. Document-level neural machine translation with hierarchical attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 2947−2954

18 Zhang B, Xiong D Y, Su J S. Neural machine translation with deep attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(1): 154−163

19 Ueffing N, Haffari G, Sarkar A. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 2007, **21**(2): 77−94

20 Bertoldi N, Federico M. Domain adaptation for statistical machine translation with monolingual resources. In: Proceedings of the 4th Workshop on Statistical Machine Translation. Athens, Greece: Association for Computational Linguistics, 2009. 182−189

21 Klementiev A, Irvine A, Callison-Burch C, Yarowsky D. Toward statistical machine translation without parallel corpora. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, 2012. 130−140

22 Zhang J J, Zong C Q. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, 2013. 1425−1434

23 Ravi S, Knight K. Deciphering foreign language. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA: Association for Computational Linguistics, 2011. 12−21

24 Dou Q, Vaswani A, Knight K. Beyond parallel data: Joint word alignment and decipherment improves machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. 557−565

25 Cheng Y, Xu W, He Z J, He W, Wu H, Sun M S, et al. Semi-supervised learning for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016. 1965−1974

26 Skorokhodov I, Rykachevskiy A, Emelyanenko D, Slotin S, Ponkratov A. Semi-supervised neural machine translation with language models. In: Proceedings of the 2018 AMTA Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018). Boston, USA: Association for Machine Translation in the Americas, 2018. 37−44

27 Artetxe M, Labaka G, Agirre E, Cho K. Unsupervised neural machine translation. In: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, Canada: OpenReview.net, 2018.

28 Lample G, Ott M, Conneau A, Denoyer L, Ranzato M A. Phrase-based and neural unsupervised machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 5039−5049

29 Burlot F, Yvon F. Using monolingual data in neural machine translation: A systematic study. In: Proceedings of the 3rd Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018. 144−155

30 Tishby N, Pereira F C, Bialek W. The information bottleneck method. arXiv: physics/0004057, 2000.

31 Eikema B, Aziz W. Auto-encoding variational neural machine translation. In: Proceedings of the 4th Workshop on Representa-

tion Learning for NLP (RepL4NLP-2019). Florence, Italy: Association for Computational Linguistics, 2019. 124−141

32 Su J S, Wu S, Xiong D Y, Lu Y J, Han X P, Zhang B. Variational recurrent neural machine translation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI, 2018. 5488−5495

33 Kingma D P, Ba L J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, USA, 2014.

34 Sennrich R, Zhang B. Revisiting low-resource neural machine translation: A case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL2019). Florence, Italy: Association for Computational Linguistics, 2019. 211−221

35 Papineni K, Roukos S, Ward T, Zhu W J. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, USA: Association for Computational Linguistics, 2002. 311−318

36 Isozaki H, Hirao T, Duh K, Sudoh K, Tsukada H. Automatic evaluation of translation quality for distant language pairs. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, USA: Association for Computational Linguistics, 2010. 944−952

37 Koehn P. Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP2004). Barcelona, Spain: Association for Computational Linguistics, 2004. 388−395

38 Zhang Z R, Liu S J, Li M, Zhou M, Chen E H. Joint training for neural machine translation models with monolingual data. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI Press, 2018. Article No. 69

于志强 昆明理工大学信息工程与自动化学院博士研究生. 主要研究方向为自然语言处理.
E-mail: yzqyt@hotmail.com
(**YU Zhi-Qiang** Ph.D. candidate at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His main research interest is natural language processing.)

余正涛 昆明理工大学信息工程与自动化学院教授. 主要研究方向为自然语言处理. 本文通信作者.
E-mail: ztyu@hotmail.com
(**YU Zheng-Tao** Professor at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His main research interest is natural language processing. Corresponding author of this paper.)

黄于欣 昆明理工大学信息工程与自动化学院博士研究生. 主要研究方向为自然语言处理.
E-mail: huangyuxin2004@163.com
(**HUANG Yu-Xin** Ph.D. candidate at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His main research interest is natural language processing.)

郭军军 昆明理工大学信息工程与自动化学院讲师. 主要研究方向为自然语言处理.
E-mail: guojjgb@163.com
(**GUO Jun-Jun** Lecturer at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. His main research interest is natural language processing.)

高盛祥 昆明理工大学信息工程与自动化学院副教授. 主要研究方向为自然语言处理.
E-mail: gaoshengxiang.yn@foxmail.com
(**GAO Sheng-Xiang** Associate professor at the Faculty of Information Engineering and Automation, Kunming University of Science and Technology. Her main research interest is natural language processing.)