Journal of Computer Applications
计算机应用, 2021, 41(6): 1652 - 1658
ISSN 1001-9081
CODEN JYIIDU
2021-06-10
http://www.joca.cn

文章编号:1001-9081(2021)06-1652-07
DOI:10.11772/j.issn.1001-9081.2020071017

融合单语语言模型的汉越伪平行语料生成

贾承勋[1,2],赖　华[1,2],余正涛[1,2*],文永华[1,2],于志强[1,2]
(1. 昆明理工大学 信息工程与自动化学院,昆明 650504;
2. 云南省人工智能重点实验室(昆明理工大学),昆明 650500)
(*通信作者电子邮箱ztyu@hotmail.com)

摘　要:神经机器翻译在资源丰富的语种上取得了良好的翻译效果,但是由于数据稀缺问题在汉语–越南语这类低资源语言对上的性能不佳。目前缓解该问题最有效的方法之一是利用现有资源生成伪平行数据。考虑到单语数据的可利用性,在回译方法的基础上,首先将利用大量单语数据训练的语言模型与神经机器翻译模型进行融合,然后在回译过程中通过语言模型融入语言特性,以此生成更规范质量更优的伪平行数据,最后将生成的语料添加到原始小规模语料中训练最终翻译模型。在汉越翻译任务上的实验结果表明,与普通的回译方法相比,通过融合语言模型生成的伪平行数据使汉越神经机器翻译的BLEU值提升了1.41个百分点。

关键词:汉越神经机器翻译;数据增强;伪平行数据;单语数据;语言模型
中图分类号:TP391　　**文献标志码**:A

Chinese-Vietnamese pseudo-parallel corpus generation based on monolingual language model

JIA Chengxun[1,2], LAI Hua[1,2], YU Zhengtao[1,2*], WEN Yonghua[1,2], YU Zhiqiang[1,2]
(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan* 650504, *China*;
2. *Yunnan Key Laboratory of Artificial Intelligence (Kunming University of Science and Technology), Kunming Yunnan* 650500, *China*)

Abstract: Neural machine translation achieves good translation results on resource-rich languages, but due to data scarcity, it performs poorly on low-resource language pairs such as Chinese-Vietnamese. At present, one of the most effective ways to alleviate this problem is to use existing resources to generate pseudo-parallel data. Considering the availability of monolingual data, based on the back-translation method, firstly the language model trained by a large amount of monolingual data was fused with the neural machine translation model. Then, the language features were integrated into the language model in the back-translation process to generate more standardized and better quality pseudo-parallel data. Finally, the generated corpus was added to the original small-scale corpus to train the final translation model. Experimental results on the Chinese-Vietnamese translation tasks show that compared with the ordinary back-translation methods, the Chinese-Vietnamese neural machine translation has the BiLingual Evaluation Understudy (BLEU) value improved by 1.41 percentage points by fusing the pseudo-parallel data generated by the language model.

Key words: Chinese-Vietnamese neural machine translation; data augmentation; pseudo-parallel data; monolingual data; language model

0　引言

神经机器翻译(Neural Machine Translation, NMT)是Sutskever等[1]提出的端到端的机器翻译方法,其训练数据越多模型性能越好,但对于资源稀缺型语言而言,可获取的双语数据十分有限,这也是导致翻译效果不佳的主要原因。

目前改善低资源语言神经机器翻译系统性能的方法有很多,其中利用现有资源扩充伪平行数据的方法是目前较为有效的方法之一。目前实现数据扩充的方法主要有四类:第一类方法是在可比语料中抽取伪平行句对[2-3],通过将源语言与目标语言映射到同一空间中,根据一定规则挑选出候选平行句对,这种方法能够有效地抽取伪平行语料,但是不容易捕捉句子特征,并且抽取到的伪平行句对噪声较大;第二类方法是基于词的替换[4-5],利用现有小规模平行句对指定的词进行规则替换得到新的伪平行句对,但是当出现单词一对多的情况

收稿日期:2020-07-13;**修回日期**:2021-01-27;**录用日期**:2021-02-01。
基金项目:国家自然科学基金资助项目(61672271,61732005,61761026,61762056,61866020);国家重点研发计划项目(2019QY1801)。
作者简介:贾承勋(1994—),男,内蒙古赤峰人,硕士,主要研究方向:机器翻译、自然语言处理;　赖华(1966—),男,广西钦州人,副教授,硕士,CCF会员,主要研究方向:智能信息处理;　余正涛(1970—),男(蒙古族),云南曲靖人,教授,博士,CCF会员,主要研究方向:自然语言处理、机器翻译;　文永华(1979—),男(白族),云南大理人,博士研究生,CCF会员,主要研究方向:机器翻译;　于志强(1983—),男,内蒙古通辽人,博士研究生,主要研究方向:机器翻译。

时效果不佳；第三类是基于枢轴语言的方法[6]，文献[7]将其整理分为系统级、语料级以及短语级三种方法，并提出通过扩大生成训练数据的规模以及优化词对齐质量的方式来提高系统的翻译性能，此方法适用于零资源语言但产生的语料质量不佳，针对此问题文献[8]将源–枢轴及枢轴–目标语言的稀有词整理为双语词典并融入到枢轴语言方法的翻译过程中，有效地提升了枢轴语言方法生成伪平行数据的质量；第四类是利用单语数据进行回译(Back-Translation，BT)[9]，通过小规模训练数据训练目标语言到源语言的翻译模型，将目标语言单语数据翻译为源语言数据，以此生成伪平行数据。

汉语–越南语是典型的低资源语言对，可获取的平行语料较少，通过数据扩充生成伪平行数据可以较好地缓解此类问题。考虑到单语数据易于获取且资源充足，但大多数现有的方法没有充分利用单语资源，因此本文针对利用单语数据生成伪平行语料的方法进行了探索研究。由于利用大量单语数据训练的语言模型可以较好地学习到语言特性，因此本文将单语语言模型与神经机器翻译模型融合，使得在伪平行数据生成过程中可以通过语言模型融入目标语言的语言特性。实验结果表明，相较于基准系统，本文方法生成的伪平行数据能有效提高汉越神经机器翻译的性能。

1　相关工作

近年来，国内外相关研究人员对单语言数据如何提升系统翻译性能进行了广泛研究，文献[10]将利用单语数据提升神经机器翻译性能的方法分为与体系结构相关的方法和与体系结构无关的方法。与体系结构相关的方法是需要神经机器翻译模型的特定结构特征或需要对体系结构进行更改；与体系结构无关方法是使用单语语料生成伪平行语料，然后将伪平行语料与平行语料混合。

目前，目标语言端的单语数据已经被证实能够极大地提升模型的翻译质量，并被广泛利用，最有效的就是文献[9]提出的回译方法，即反向翻译，使用预先训练的机器翻译系统翻译目标语言的单语数据，从而生成大量的伪双语数据，并将这些伪双语数据添加到原始数据中进行源语言到目标语言翻译模型的训练，但是其翻译生成的句子会存在许多错误，从而影响源语言到目标语言的翻译模型训练。文献[11]利用不同性能的翻译模型通过回译生成质量不同的伪平行数据，研究伪平行数据质量对性能提升的影响，证明了伪平行数据的质量越好对模型性能的提升也会越高，但其只是通过不同模型改变质量，并没有对如何提升伪平行数据质量进行研究。在文献[12]和文献[13]的研究中表明，通过正向翻译(Forward Translation，FT)获得的人工平行语料也可以证明是有利的，但是提升效果相比反向翻译较差。文献[14]对伪平行数据有效性的上界进行了探索，伪平行数据与原始数据的比例不超过8∶1，就不会降低系统的性能。文献[15]提出了一种将源语言单语数据和目标语言单语数据进行联合训练的方法，通过同时训练源到目标和目标到源的翻译模型，在训练过程中同时利用正向和反向生成的伪平行数据对两个模型进行迭代训练。文献[16]提出利用源语言和目标语言的大规模单语数据，通过正向翻译和反向翻译生成伪平行数据，实验结果表明只使用源端或目标端的单语数据生成更多的伪平行数据，对模型的提升效果不会随着数据量的增加而增加。

以上方法均是利用单语数据生成伪平行数据提升神经机器翻译的性能，但对低资源语言神经机器翻译性能的提升仍然有限。汉语和越南语都是独立派系的语言且汉越双语训练数据稀缺，考虑到伪平行数据的数量对系统性能的提升有限，而语言模型容易通过训练得到，因此本文在伪平行数据的生成过程中，将利用大量单语数据训练得到目标语言语言模型融合到神经机器翻译模型中，融合目标语言模型的预期效果是通过语言模型在伪平行数据的生成中融入语言特性，帮助生成语法正确的句子，使得到的伪平行数据更加规范，从而提高伪平行数据的质量。由于正向翻译和反向翻译生成的数据均可以提升系统的性能，同时为了充分证明融合单语语言模型方法的有效性，本文在正向和反向上都生成了汉越伪平行数据，其中正向翻译中融合的是越南语越南源语言模型，反向翻译中融合的是汉语语言模型。由于生成的伪平行语料中包含部分噪声，因此本文对生成的伪平行语料利用汉语和越南语语言模型对其进行质量筛选，将最后得到的伪平行数据与原始数据一起训练最终汉越神经机器翻译模型。

2　单语数据生成伪平行数据方法

2.1　伪平行数据生成框架

目前已知正向翻译和回译生成的伪平行数据对系统性能均有提升，因此在两个方向上生成伪平行数据，并对其对翻译系统提升的效果进行了实验对比。对于语言模型与翻译模型的融合，本文进行了两种融合方法的实验，分别称为基于独立训练的语言模型融合和基于合并训练的语言模型融合。语言模型的选择上采用可以处理任意长度输入序列的循环神经网络语言模型(Recurrent Neural Network Language Model，RNNLM)[17]，整体流程如图1所示。

语言模型可以看作是对一个句子存在概率的评估，通过将语言模型融合后对单语数据进行翻译生成伪平行数据，在伪平行数据生成过程中可以结合目标语言的语言特性。图1中对正向翻译和反向翻译生成的伪平行数据进行筛选所用的汉越语言模型是相同的，均是利用相同的单语数据训练得到的语言模型。以下将从生成伪平行数据的方式、RNNLM和NMT独立训练融合、RNNLM和NMT合并训练融合以及语言模型困惑度语料筛选几个方面对本文利用单语数据实现语料扩充的方法进行说明。

2.2　融合语言模型的伪平行数据生成方法

反向翻译方法被证实是一种非常有效并且能较好地提升翻译系统性能的方法，该方法有效地利用了目标语言的单语数据。反向翻译具体流程如图1所示，首先使用小规模汉越双语语料 $D=\{(x^{(n)},y^{(n)})\}_{n=1}^{N}$ 训练一个越汉翻译模型 $M_{y\to x}$，然后将此翻译模型与外部语言模型进行融合，同时将越南语单语数据 $Y=\{y^{(t)}\}_{t=1}^{T}$ 翻译为汉语数据 $X'=\{x^{(t)}\}_{t=1}^{T}$，在翻译期间通过融合的语言模型对翻译的数据结合越南语的语言特性使其规范化，以此构成反向翻译生成汉越伪平行数据 $\tilde{D}_{y\to x}=\{x'^{(t)},y^{(t)}\}_{t=1}^{T}$，然后将生成的伪平行数据经过语言模型筛选后与原始数据一起训练汉越神经机器翻译模型。

正向翻译方法生成的伪平行数据对系统的提升效果比反向翻译的略差一些，但对系统的翻译性能仍有提升。正向翻译具体流程如图1所示，使用小规模双语语料 D，训练汉到越的翻译模型，然后用此模型将汉语单语数据 $X=\{x^{(t)}\}_{t=1}^{T}$ 翻译为越南语数据，在翻译过程中本文将外部单语语言模型融合进来，通过语言模型将越南语语言特性结合进来，生成通过正向翻译的汉越伪平行数据 $\tilde{D}_{y\to x}=\{x'^{(t)},y^{(t)}\}_{t=1}^{T}$，最后将生成的伪平行数据与原始数据一起训练汉越神经机器翻译模型。

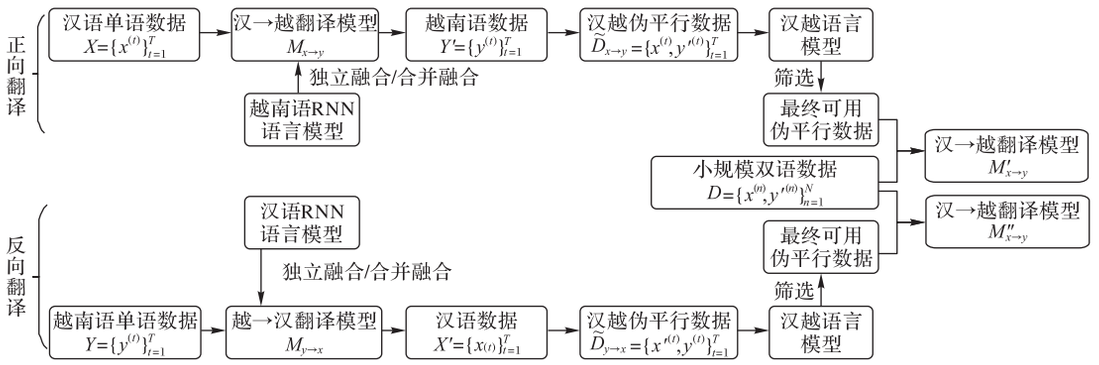

图1 伪平行数据生成框架整体流程

Fig. 1 Overall flowchart of pseudo parallel data generation framework

2.2.1 基于独立训练的语言模型融合

语言模型训练方便并且可以学习到较好的语言特性，对翻译性能可以有很好的提升效果，因此本文探索了语言模型独立于翻译模型的融合方式，因为模型间相互独立所以对语言模型的架构没有限制，本文可以选择基于 n 元语法（n-gram）的前馈语言模型[18]或是基于循环神经网络的语言模型，由于循环神经网络语言模型通过使用词向量（Word Embedding）作为输入能够在一定程度上缓解数据稀疏问题，并且循环结构的引入可以对长距离信息进行有效建模，获得更好的语言模型性能，因此本文方法选用循环神经网络语言模型进行实验。

RNNLM与NMT独立训练融合，是对NMT与RNNLM分别进行训练，最后在模型softmax层输出概率进行拼接融合的方法。首先利用大量越南语单语语料对语言模型进行预训练，同时利用现有数据训练一个汉越神经机器翻译模型，然后在神经机器翻译模型每一时间步长预测下一个单词的时候，将NMT的概率分布与RNN语言模型的概率分布进行加权合并，以包含注意力机制（Attention Mechanism）[19]的RNNsearch模型[20]为例，模型融合后在 t 时刻下翻译流程如图2所示。

在神经机器翻译模型和RNN语言模型的每个时间步长，翻译模型和语言模型都会根据前一时刻预测的单词对建议下一个可能的单词进行概率预测，然后将NMT预测的概率 $P_{\mathrm{NMT}}(y_t|x)$ 与语言模型预测的概率 $P_{\mathrm{LM}}(y_t)$ 乘以超参数 λ 相加，最后概率最高的单词被选为序列中的下一个单词 y_t，式（1）为NMT模型在 t 时刻预测的单词概率，RNNLM与NMT独立训练融合的思想便是将 y_t 的概率预测从式（1）修改成式（2）：

$$\log y_t = \arg\max_{y} \log P_{\mathrm{NMT}}(y|x) \tag{1}$$

$$\log y_t = \arg\max_{y} \log P_{\mathrm{NMT}}(y|x) + \lambda \log P_{\mathrm{LM}}(y) \tag{2}$$

其中：x 为源语言词，y 为目标语言词，y_{t-1} 为前一时刻预测的目标语言单词；λ 为超参数，作为语言模型译句的概率分布的权重，此方法需要对验证数据进行额外的微调，以控制语言模型的影响。为了使单词序列获得得更加准确，神经机器翻译模型中解码器应用集束搜索（Beam Search），选择beam size=3，即选择概率最大的产生3个最可能的序列，直到预测结束为止，然后选择概率最高的序列。

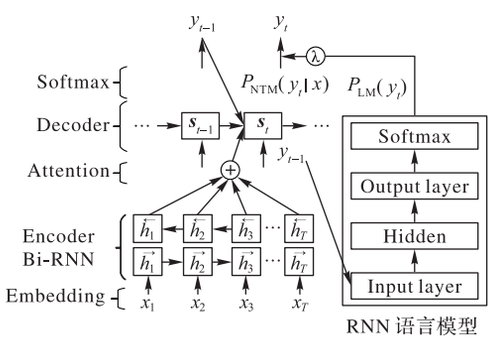

图2 t 时刻独立训练融合方法的翻译流程

Fig. 2 Translation flowchart of independent training fusion method at t time

通过此融合后的模型，利用正向翻译和反向翻译方法生成伪平行数据，与原始数据混合后再进行汉越神经机器翻译模型训练。

2.2.2 基于合并训练的语言模型融合

在训练过程中，考虑到更深的融合可以更好地融合语言特性，因此本文对语言建模集成到神经机器翻译模型体系结构中的方法进行了实验。合并训练融合的好处是，考虑到神经网络体系结构的特征，本文可以更有效地利用循环神经网络语言模型。合并训练融合直接将基于循环神经网络的语言模型的隐状态和神经机器翻译模型解码器的隐状态合并在一起，然后用此合并的隐状态预测最终翻译概率，合并训练融合方法的结构流程如图3所示。

如图3所示，与原始神经机器翻译模型不同，最终的隐藏层除了NMT的隐藏状态外，还将循环神经网络语言模型的隐藏状态作为输入。其中 y_t 为 t 时刻预测的目标语言单词，$\boldsymbol{C}_t$ 为 t 时刻的上下文向量，$\boldsymbol{s}_t^{\mathrm{LM}}$ 为语言模型 t 时刻的隐藏层状态，$\boldsymbol{s}_t^{\mathrm{NMT}}$ 为神经机器翻译模型 t 时刻解码器的隐状态，在每一时间步长中语言模型和翻译模型的隐状态还将嵌入前一时刻预测的单词序列。因此影响整体性能的因素为上下文向量 $\boldsymbol{C}_t$、前一时刻的单词序列 y_{t-1} 以及隐向量状态 $\boldsymbol{s}^{\mathrm{NMT}}$ 和 $\boldsymbol{s}^{\mathrm{LM}}$，$t$ 时刻预测单词概率如式（3）所示：

$$\log y_t = \arg\max \log P(y|x, y < t) \tag{3}$$

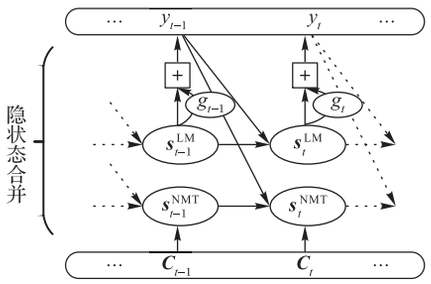

图3　隐状态合并训练流程

Fig. 3　Flowchart of hidden state merge training

为了平衡语言模型对神经机器翻译模型的影响,用一个控制器网络g_t在每一步计算中调整语言模型隐状态和解码器隐状态之间的权重,并根据训练数据对模型的隐藏输出和控制器机制参数进行微调,具体过程如式(4)~(5)所示:

$$P(y_t|y < t, x) \propto \exp(y_t^{\mathrm{T}}(\boldsymbol{W}_o f_o(\boldsymbol{s}_t^{\mathrm{LM}}, \boldsymbol{s}_t^{\mathrm{NMT}}, y_{t-1}, \boldsymbol{C}_t) + b_o)) \tag{4}$$

其中:$\boldsymbol{W}_o$是学习得到的权重矩阵;f_o是具有双向最大非线性输出的单层前馈神经网络;b_o为偏差。

$$g_t = \sigma(\boldsymbol{v}_g^{\mathrm{T}}\boldsymbol{s}_t^{\mathrm{LM}} + b_g) f_o \tag{5}$$

其中:σ是logistic函数;$\boldsymbol{v}_g^{\mathrm{T}}$和$b_g$是学习参数。通过将控制器的输出与语言模型的隐状态相乘,使解码器可以充分利用NMT的信号,而控制器则控制语言模型信号的权重。同时为了使语言模型所学到的越南语特性不被覆盖,在训练过程中,只对用于参数化输出的参数进行调整。

在汉到越的情况下,当没有与中文单词相对应的越南语单词时,在这种情况下语言模型可以提供更多信息,同时如果要翻译单词为名词时,则最好忽略来自语言模型隐藏层的信号,因为名词对后续单词概率预测的影响较大,这可能会影响解码器选择正确的翻译。

2.3　基于语言模型困惑度的数据筛选

在统计机器翻译中语料质量评价的方法有很多,路琦等[21]对其训练语料质量的筛选方法进行了详细的研究。对于神经机器翻译伪平行数据的筛选,由于语言模型计算句子的困惑度(perplexity)实现方便且准确度高,同时语言模型的困惑度评价可以评判句子中单词序列出现的合理性,可以对句子的流畅度进行评判,因此本文选择此方式来过滤低质量的伪平行语料。困惑度的评判标准是越小句对的合理性越高,句子的流畅度也越好。基于语言模型困惑度的汉越伪平行数据筛选流程如图4所示。

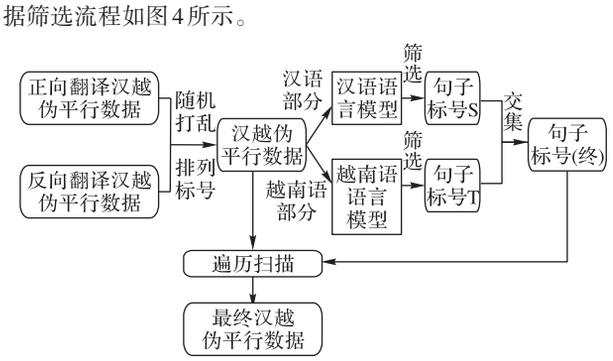

图4　伪平行数据筛选流程

Fig. 4　Flowchart of pseudo-parallel data screening

利用语言模型筛选数据的特点在于首先对生成的伪平行数据进行排序标号,然后同时利用汉语语言模型和越南语语言模型对伪平行数据中各自语言部分进行困惑度评判,通过过滤得到困惑度小于阈值的句子序号,然后根据句对的序号排列取其交集得到最终符号条件的句对序号,最后在原始伪平行数据中遍历扫描,得到最终实验可用的伪平行数据。

3　实验与结果分析

3.1　实验设置

为验证融合单语语言模型生成的汉越伪平行数据的有效性,本文分别在正向翻译和反向翻译上生成伪平行数据,并与原始数据结合训练汉越神经机器翻译模型。同时为了验证融合单语语言模型生成的数据质量要比原始模型生成的伪平行数据质量高,可以使模型获得更好的性能提升,本文分别对不同的伪平行数据对系统性能提升的影响进行对比分析。考虑到生成伪平行数据所用单语数据与训练语言模型所用单语数据的相关性对最终模型性能的影响,本文还对来自不同领域的单语数据生成的伪平行语料对系统性能的影响进行了实验对比。

实验中所用汉越双语语料是通过网络爬取并进行数据清洗后得到的160 × 10^3平行句对,并分为训练集、验证集与测试集,其中验证集、测试集为在160×10^3平行句对中随机抽取的2 × 10^3个句对。本文总共收集汉语单语语料和越南语单语语料各3 200 × 10^3,其中:3 000 × 10^3用来训练语言模型;500×10^3来自教育领域语料库QCRI;2 500 × 10^3来自维基百科(Wikipedia)20191201整理的数据集;余下200 × 10^3用来进行伪平行数据生成,QCRI和Wikipedia语料各100 × 10^3。对单语数据进行随机打乱后进行实验,实验所用数据如表1和表2所示。

表1　实验用双语数据

Tab. 1　Experimental bilingual data

数据集类型	汉越数据集大小/10^3
训练集	156
测试集	2
验证集	2

表2　单语数据利用情况

Tab. 2　Monolingual data utilization

领域	来源	正向翻译/10^3		反向翻译/10^3	
		RNNLM(越)	翻译数据(汉)	RNNLM(汉)	翻译数据(越)
教育	QCRI	500	100	500	100
综合	Wikipedia	2 500	100	2 500	100
合并		3 000	200	3 000	200

在进行实验前需要先对数据进行预处理,首先对训练数据进行tokenization处理,并将句子长度在50个词以上的句对过滤。实验中使用的神经机器翻译模型是RNNsearch和谷歌(Google)开源模型Transformer,使用的语言模型是基于循环神经网络的语言模型(RNNLM)。所有实验均使用大小为156×10^3的双语平行语料作为训练集,词表大小均设置为30×10^3,实验均在单卡GPU服务器上进行,为防止出现过拟合现象,在多次实验调整后将损失值dropout设置为0.1,批值

(batch size)为64,隐层单元(hidden units)为512,训练步长(train steps)为200×10³,使用BLEU(BiLingual Evaluation Understudy)值作为评测指标。

3.2 结果分析

实验中比较了RNNsearch模型和Transformer模型与RNNLM融合生成的伪平行数据对汉越神经机器翻译性能的影响,同时也对比了不同语言模型融合方式生成的伪平行数据对汉越神经机器翻译性能的影响。实验中baseline为仅利用原始数据训练得到的模型效果,最终模型翻译方向均为汉到越,生成的伪平行数据通过语言模型筛选在正向和反向翻译中分别过滤了5 982和8 073个句对,通过正向翻译方法扩展了194 × 10³伪平行数据,通过反向翻译方法扩展了192 × 10³可用伪平行数据。为保证实验结果的可靠性,每组结果的BLEU值都是利用相同测试集进行实验得到的结果,如表3所示。

表3 添加伪平行数据后的BLEU值

Tab. 3 BLEU value after adding pseudo-parallel data

方法	伪平行语料规模/10³	总语料规模/10³	BLEU值/%			
			RNNsearch		Transformer	
			独立	合并	独立	合并
baseline	—	156	16.84		21.28	
正向翻译	194	156+194	17.21	17.62	21.68	22.06
反向翻译	192	156+192	17.54	17.99	22.11	22.54
正+反	194+192	156+386	**17.69**	**18.19**	**22.26**	**22.69**

实验结果表明,增加伪平行数据后可以提升汉越神经机器翻译模型的翻译性能,并且通过基于合并训练融合生成的伪平行语料对翻译性能的提升效果要优于基于独立训练融合的效果,同时相较于独立训练融合方法BLUE值约平均提升了0.45个百分点。正向翻译方法生成的伪平行语料提升效果相较于反向翻译方法要略低一些,这是因为反向翻译生成的伪平行数据中越南语部分是真实语句,汉语部分为翻译生成的语句,而正向翻译刚好相反,这使得模型在进行训练时无法较为准确地获取越南语语言信息,所以反向翻译比正向翻译生成的伪平行数据对系统的提升效果要好。最后将正向和反向生成的伪平行数据合并,进一步增加了伪平行数据的数量,在Transformer模型中的BLEU值相较于baseline最高提升了1.41个百分点。

在实验中,RNNsearch模型的效果较差,这是因为基于RNN的翻译模型在训练过程中由于线性序列依赖特性很难具备高效的并行计算能力,并且编码器产生固定长度的源语言上下文向量,这种方式无法充分地利用上下文关系,而Transformer模型的编码器层是由6个encoder堆叠而成,解码器也一样。每个encoder包含两层,一个self-attention层和一个前馈神经网络,self-attention能帮助当前节点不仅仅只关注当前的词,同时能更好地获取上下文的语义信息;decoder也包含这两层网络,并在这两层中间还有一个attention层,帮助当前节点获取到当前需要关注的重点内容,所以Transformer可以更好地利用上下文信息并且充分的利用数据训练翻译模型。

为了验证融合单语语言模型方法生成的伪平行数据质量相对较好,在此对不同的伪平行数据对系统性能提升的影响进行对比分析。在RNNsearch和Transformer模型下,对比无语言模型融合与融合语言模型生成的伪平行数据对最终翻译模型性能提升的效果,其中伪平行语料规模均固定为200×10³,结果如表4所示。

从表4可以看出,基于独立训练融合生成的伪平行数据与无语言模型(无LM)生成的伪平行数据对系统性能的提升相近,影响不大,而通过基于合并训练融合生成的伪平行数据相对无语言模型生成的伪平行数据对系统性能提升较高,这是因为伪平行数据的质量得到了提高,可以进一步提升模型的翻译效果。

表4 不同伪平行语料质量对BLEU值提升的影响 单位:%

Tab. 4 Impact of different pseudo-parallel corpus qualities on BLEU value improvement unit:%

模型	baseline	正向翻译			反向翻译		
		无LM	独立	合并	无LM	独立	合并
RNNsearch	**16.84**	17.13	17.21	**17.62**	17.49	17.54	**17.99**
Transformer	**21.28**	21.57	21.68	**22.06**	22.03	22.11	**22.54**

为了验证使用与训练语言模型来自不同领域的单语语料生成的伪平行语料对模型性能提升的影响,本文在汉语-越南语翻译方向上,利用基于合并训练的语言模型融合方式,通过反向翻译方法利用越南语单语数据生成伪平行数据。其中训练语言模型的数据来自维基百科的单语语料,将生成伪平行语料的单语语料分为4种不同的组成,记为情况1~4,分别对应单语语料完全来自Wikipedia(100%)、75%与语言模型的领域相同余下部分为教育领域语料、50%相同和领域完全不同(0%),实验结果如表5所示。

表5 不同领域单语数据BLEU值对比 单位:%

Tab. 5 BLEU value comparison of different domain monolingual data unit:%

模型	情况1	情况2	情况3	情况4
RNNsearch	**18.12**	18.05	17.99	17.83
Transformer	**22.65**	22.61	22.54	22.41

从表5可以看出,当训练语言模型与翻译利用的单语数据领域相似越多,伪平行数据对最终翻译模型BLEU值的提升也会越高。

3.3 译文对比分析

以正向翻译(汉到越)生成的伪平行数据为例,对比分析融入循环神经网络语言模型后生成伪平行数据的质量影响,将汉语通过本文方法翻译为越南语,例句1"红色与蓝色混合变成了紫色。"和例句2"将来我一定会感谢那些帮助过我的人。"的不同方法翻译对比结果如图5~6所示。

通过对比不同方式生成的伪平行句对可以看出,Transformer模型生成的伪平行数据质量要优于RNNsearch模型,主要原因是Transformer模型可以更好地结合上下文信息,并且对于部分词的翻译更为准确,如"与蓝色混合"译文为图7(a);而RNNsearch的译文为图7(b),存在明显的句法错误和词的翻译问题。同时可以看出,通过基于合并训练的语言模型融合方式翻译得到的越南语译文质量比基于独立训练融合方式更优,如"那些帮助过我的人"的正确译文为图7(c),合并融合方式翻译得到的越南语译文更加符合越南语语言特性,而独立训练融合方式效果相对较弱。

Màu đỏ(红色) pha trộn(混合) với(与) màu xanh lam(蓝色) thành(变成) màu tím(紫色).
(a) 参考译文
Màu đỏ(红色) và(与) xanh(绿色) được trộn(混在一起) vào(在) tím(紫色的).
(b) RNNsearch
Màu đỏ(红色) và(与) màu xanh(绿色) được trộn(混在一起) vào(在) màu tím(紫色).
(c) RNNsearch+独立
Màu đỏ(红色) được trộn(混在一起) với(与) màu xanh lam(蓝色) vào(在) màu tím(紫色).
(d) RNNsearch+合并
Màu đỏ(红色) với(与) màu xanh lam(蓝色) pha trộn(混合) thành(变成) màu tím(紫色).
(e) Transformer+独立
Màu đỏ(红色) pha trộn(混合) với(与) màu xanh lam(蓝色) thành(变成) màu tím(紫色).
(f) Transformer+合并

图5　不同方法翻译结果对比(例句1)

Fig. 5　Comparison of translation results of different methods (example 1)

Sau này(将来) tôi(我) nhất định(一定) sẽ(会) cảm ơn(感谢) những(那些) người(人) từng(曾经) giúp đỡ(帮助) tôi(我).
(a) 参考译文
Trong tương lai(将来) tôi(我) chắc chắn(当然) sẽ(会) cảm ơn(感谢) những(那些) đã(是) giúp(帮忙) tôi(我) người(人) .
(b) RNNsearch
Trong tương lai(将来) tôi(我) chắc chắn(当然) sẽ(会) cảm ơn(感谢) những(那些) đã(是) giúp đỡ(帮助) tôi(我) người(人) .
(c) RNNsearch+独立
Trong tương lai(将来) tôi(我) chắc chắn(当然) sẽ(会) cảm ơn(感谢) những(那些) người(人) từng(曾经) giúp đỡ(帮助) tôi(我) .
(d) RNNsearch+合并
Sau này(将来) tôi(我) chắc chắn(当然) sẽ(会) cảm ơn(感谢) những(那些) từng(曾经) giúp đỡ(帮助) tôi(我) người(的人).
(e) Transformer+独立
Sau này(将来) tôi(我) nhất định(一定) sẽ(会) cảm ơn(感谢) những(那些) người(人) từng(曾经) giúp đỡ(帮助) tôi(我) .
(f) Transformer+合并

图6　不同方法翻译结果对比(例句2)

Fig. 6　Comparison of translation results of different methods (example 2)

pha trộn(混合) với(与) màu xanh lam(蓝色)
(a) “与蓝色混合”的译文(Transformer)
và(与) xanh(绿色) được trộn(混在一起)
(b) “与蓝色混合”的译文(RNNsearch)
những(那些)người(人) từng(曾经) giúp đỡ(帮助) tôi(我)
(c) “那些帮助过我的人”的正确译文

图7　部分词的译文示例

Fig. 7　Translation examples of some words

4　结语

本文针对汉越神经机器翻译数据稀缺问题,充分利用单语数据资源,提出了利用单语数据在正向和反向两个方向上生成伪平行数据的过程中将循环神经网络语言模型融合到神经机器翻译模型中的方法,通过语言模型结合语言特性,从而提升了伪平行数据的质量。实验结果表明,本文方法与单一的正向翻译和反向翻译生成方法相比,可以在汉越神经机器翻译中通过提升伪平行数据质量从而更好地提升翻译系统的性能。在未来工作中,我们会探索单语数据的选择以及伪平行数据与原始数据的权重比对系统翻译性能的影响。

参考文献 (References)

[1] SUTSKEVER I, VINYALS O, LE Q V, et al. Sequence to sequence learning with neural networks [C]// Proceedings of the 2014 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 3104-3112.

[2] MARIE B, FUJITA A. Efficient extraction of pseudo-parallel sentences from raw monolingual data using word embeddings [C]// Proceedings of the 2017 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 392-398.

[3] ABDUL RAUF S, SCHWENK H. Parallel sentence generation from comparable corpora for improved SMT [J]. Machine Translation, 2011, 25(4): 341-375.

[4] FADAEE M, BISAZZA A, MONZ C. Data augmentation for low-resource neural machine translation [C]// Proceedings of the 2017 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 567-573.

[5] 蔡子龙,杨明明,熊德意. 基于数据增强技术的神经机器翻译[J]. 中文信息学报,2018,32(7):30-36. (CAI Z L, YANG M M, XIONG D Y. Data augmentation for neural machine translation [J]. Journal of Chinese Information Processing, 2018, 32(7): 30-36.)

[6] ZAHABI S T, BAKHSHAEI S, KHADIVI S. Using context vectors in improving a machine translation system with bridge language [C]// Proceedings of the 2013 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2013: 318-322.

[7] 李强,王强,肖桐,等. 稀缺资源机器翻译中改进的语料级和短语级中间语方法研究[J]. 计算机学报,2017,40(4):925-938. (LI Q, WANG Q, XIAO T, et al. Research on improved corpus-level

and phrase-level pivot language based methods in low- resource machine translation [J]. Chinese Journal of Computers, 2017, 40(4): 925-938.)

[8] 贾承勋,赖华,余正涛,等. 基于枢轴语言的汉越神经机器翻译伪平行语料生成[J]. 计算机工程与科学,2021,43(3):543-550. (JIA C X, LAI H, YU Z T, et al. Pseudo-parallel corpus generation for Chinese-Vietnamese neural machine translation based on pivot language [J]. Computer Engineering and Science, 2021, 43(3): 543-550.)

[9] SENNRICH R, HADDOW B, BIRCH A, et al. Improving neural machine translation models with monolingual data [C]// Proceedings of the 2016 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2016: 86-96.

[10] GIBADULLIN I, VALEEV A, KHUSAINOVA A, et al. A survey of methods to leverage monolingual data in low-resource neural machine translation [EB/OL]. [2020-06-20]. https://arxiv. org/pdf/1910. 00373. pdf.

[11] BURLOT F, YVON F. Using monolingual data in neural machine translation: a systematic study [C]// Proceedings of the 2018 3rd Conference on Machine Translation. Stroudsburg: ACL, 2018: 144-155.

[12] PARK J, SONG J, YOON S. Building a neural machine translation system using only synthetic parallel data [EB/OL]. [2020-06-20]. https://arxiv. org/pdf/1704. 00253. pdf.

[13] CREGO J, SENELLART J. Neural machine translation from simplified translations [EB/OL]. [2020-06-20]. http://arxiv. org/pdf/1612. 06139. pdf.

[14] STAHLBERG F, CROSS J, STOYANOV V. Simple fusion: return of the language model [C]// Proceedings of the 2018 3rd Conference on Machine Translation. Stroudsburg: ACL, 2018: 204-211.

[15] ZHANG Z, LIU S, LI M, et al. Joint training for neural machine translation models with monolingual data [C]// Proceedings of the 2018 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 555-562.

[16] WU L, WANG Y, XIA Y, et al. Exploiting monolingual data at scale for neural machine translation [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2019: 4207-4216.

[17] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model [C]// Proceedings of the 2010 11th Annual Conference of the International Speech Communication Association. Piscataway: IEEE, 2010: 1045-1048.

[18] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.

[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 2017 31st International Conference on Neural Information Proceeding Systems. Red Hook: Curran Associates Inc., 2017:6000-6010.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2020-06-20]. https://arxiv. org/pdf/1409. 0473. pdf.

[21] 路琦,张傲,刘金花,等. 面向统计机器翻译的训练语料质量评价方法研究及应用[C]//第六届全国青年计算语言学会议. 北京:中国中文信息学会,2012:264-275. (LU Q, ZHANG A, LIU J H, et al. Research and application of training corpus quality evaluation method for statistical machine translation [C]// Proceedings of the 2012 6th Youth Conference on Computational Linguistics. Beijing: Chinese Information Processing Society of China, 2012: 264-275.)

This work is partially supported by the National Natural Science Foundation of China (61672271, 61732005, 61761026, 61762056, 61866020), the National Key Research and Development Program of China (2019QY1801).

JIA Chengxun, born in 1994, M. S. His research interests include machine translation, natural language processing.

LAI Hua, born in 1966, M. S., associate professor. His research interests include intelligent information processing.

YU Zhengtao, born in 1970, Ph. D., professor. His research interests include natural language processing, machine translation.

WEN Yonghua, born in 1979, Ph. D. candidate. His research interests include machine translation.

YU Zhiqiang, born in 1983, Ph. D. candidate. His research interests include machine translation.