文章编号:1671-6833(2025)06-0008-07

基于伪孪生网络的无监督学习多语言神经机器翻译方法

都力铭[1], 屈 丹[1,2], 张传财[2], 席阳丽[1]

(1.郑州大学 网络空间安全学院,河南 郑州 450002;2.网络空间部队信息工程大学 先进计算与智能工程(国家级)实验室,河南 郑州 450001)

摘 要:无监督神经机器翻译采用单语数据进行训练时会产生大量噪音信息,使得机器翻译模型在训练迭代过程中的误差不断积累,影响翻译效果。针对此问题,在跨语言预训练模型(XLM)的基础上,提出了一种基于伪孪生网络的无监督神经机器翻译方法。该方法将模型编码器分为两个模块,其中伪孪生网络部分引入了一种噪声过滤门机制,利用其对编码过程中的噪音特征进行过滤,使得模型能够更好地学习源语言和目标语言之间的映射关系。实验结果表明:在英语同德语、法语、罗马尼亚语3种语言之间的交互翻译任务中,所提方法相较于基线系统平均提升了3.5百分点,证明了其翻译效果的优越性,并使用消融实验对该模型各组件进行了有效性验证,同时在德译英翻译任务中模拟了该方法在不同噪声条件下的性能测试,表现出较好的抗噪性。

关键词:无监督机器翻译; 伪孪生网络; 单语数据; 噪声过滤门机制; 跨语言预训练模型

中图分类号:TP391;TP18 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.03.008

端到端的编码器-解码器[1]结构的提出,给传统机器翻译提供了优越的技术支持。近年来,一些基于深度学习的机器翻译模型在处理不同翻译任务时表现出色,如循环神经网络[2]、长短时记忆网络[3]和 Transformer[4]等。然而,这些方法中所提出的神经网络模型都依赖于大规模的平行语料数据。对于某些语言和领域,大规模平行语料数据是无法获取的,有些语言甚至没有可用的平行语料库。因此,为减少神经机器翻译模型对于平行语料数据的高度依赖,Lample 等[5]提出了一种使用单语语料的无监督神经机器翻译方法。然而,该方法并不完美,由于没有真正的双语数据,采用无监督的方法通常需要反向翻译等方法来生成伪双语数据,从而将单语数据用于模型训练,但是这样的训练数据往往会包含大量噪音信息和错误信息,这些噪音信息使得模型在反向翻译的迭代过程中出现的错误不断堆叠和积累,从而影响到神经机器翻译模型的性能,使翻译效果不好。

为了解决上述问题,本文借鉴 Li 等[6]提出的孪生网络的思想,基于 Conneau 等[7]的跨语言预训练模型(cross-lingual language model, XLM)提出了一种基于伪孪生网络(pseudo-siamese unsupervised machine translation, PSUMT)的无监督神经机器翻译方法。通过编码器的两个网络进行参数共享,设计出一种带期望门的注意力机制,对编码过程中的噪声进行过滤,指导模型进行文本编码,辅助模型学习双语之间的语义等方面的对应关系,为解码器提供更符合上下文的语义特征,从而提高模型无监督翻译质量。本文所提模型是在英语和法语、德语和罗马尼亚语之间的翻译任务上进行测试,实验结果表明,本文所提出的无监督神经机器翻译模型有较好的翻译效果。

1 相关工作

1.1 无监督神经机器翻译

在早期的研究工作中,缺乏平行语料数据是影

收稿日期:2025-01-12;**修订日期**:2025-02-17

基金项目:国家自然科学基金资助项目(62171470);河南省中原科技创新领军人才项目(234200510019);河南省自然科学基金项目(232300421240)

通信作者:屈丹(1974—),女,吉林九台人,网络空间部队信息工程大学教授,博士,博士生导师,主要从事语音识别、智能信息处理、机器学习等方面的研究,E-mail:qudanqudan@163.com。

引用本文:都力铭,屈丹,张传财,等. 基于伪孪生网络的无监督学习多语言神经机器翻译方法[J]. 郑州大学学报(工学版),2025,46(6):8-14.(DU L M, QU D, ZHANG C C, et al. Unsupervised learning multilingual neural machine translation based on pseudo-siamese network[J]. Journal of Zhengzhou University (Engineering Science),2025,46(6):8-14.)

响神经机器翻译性能的因素之一[8]。为了降低对大规模平行语料库的依赖,有研究者在无监督的词嵌入工作中提出了解决方案[9],其核心思想是让语义相似的词在嵌入空间中彼此更接近,采用对抗训练的方式,让模型学习从源语言到目标语言的线性映射。Vincent 等[10]提出了去噪自编码器模型,该模型改变了编码器直接复制输入的方式,训练时在输入端添加噪声序列,提高了模型的训练效率。传统无监督机器翻译方法可分为三部分,分别是初始化、去噪自编码器训练和反向翻译,这三部分是无监督机器翻译最为核心的内容。这里假设有语言 L_1,相应单语数据集为 X;语言 L_2,相应单语数据集为 Y。整体的训练过程如图 1 所示。

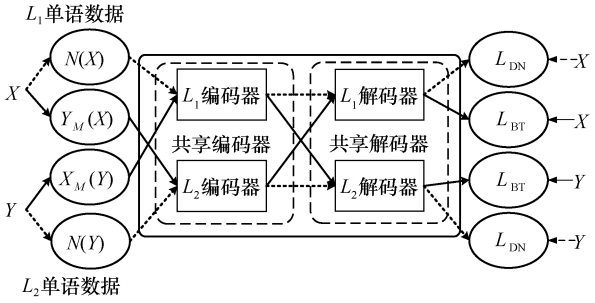

图 1　无监督机器翻译示意图

Figure 1　Schematic diagram of unsupervised machine translation

图 1 中,L_{DN} 和 L_{BT} 分别为去噪自动编码和反向翻译的目标函数,$N(X)$ 和 $N(Y)$ 均为加噪后的句子,$Y_M(X)$ 和 $X_M(Y)$ 均为由模型生成的译文,用来构成伪平行语句对。具体的迭代过程包括语言 L_1 和语言 L_2 自身的去噪自编码器训练,以及二者之间互相交替进行的反向翻译训练[11]。在训练过程中,反向翻译实时地在每个批次的语句上进行,而非对所有语料库直接进行反向翻译。无监督神经机器翻译模型最后的训练目标是通过最小化去噪自编码器和反向翻译的损失函数之和来完成的。

1.2　跨语言预训练模型

在神经网络的初始化过程中引入预训练方法,使得表征向量能够更好地映射到潜在空间[12]。这一方法虽在翻译效果上有提升,但无法很好地获取上下文之间的语义关系。Melamud 等[13]在词嵌入过程中引入双向长短期记忆网络,有效地在词向量当中融入了上下文语义信息。Devlin 等[14]提出了一种叫作 BERT(bidirectional encoder representations from Transformers)的预训练模型,该模型基于 Transformer,只采用了编码器的结构,在大量的单语数据集上进行训练,效果有着显著提升,而本文采用的基线系统 XLM 跨语言预训练模型就是基于 BERT 建模的。目前研究者们在跨语言预训练方面做了许多工作,后续提出了许多模型的变体,如 MASS[15]、XLM-R[16]、mBART[17]、DeltaLM[18]等。

1.3　孪生网络

孪生网络(siamese network)最早由 Bromley 等[19]提出,其核心思想是将原始数据同时输入到两个完全相同的神经网络中,这两个网络拥有相同的权重并且参数共享,孪生网络结构如图 2 所示。

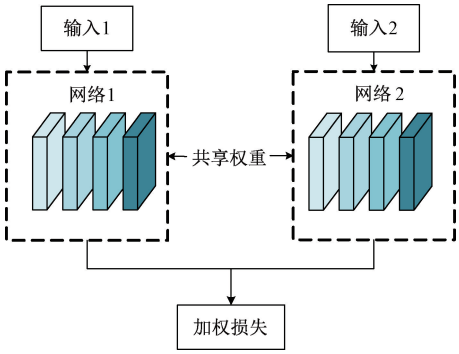

图 2　孪生网络结构图

Figure 2　Diagram of the siamese network structure

通过学习输入数据在这两个网络中的表示,进而可以计算出两个输入样本之间的相似度,以达到优化模型的目的。若将孪生网络结构中的网络 1 和网络 2 改变成两个不相同的网络,且彼此之间的参数和权重不相互共享,那么该网络架构称为伪孪生网络架构,可便于提取输入到编码器中的不同语言文本序列特征。本文所提出的模型就采用了这样的架构,将于第 2 节详细介绍。

2　基于伪孪生网络的无监督机器翻译模型

本文以 XLM 为基准模型,在此基础上对模型进行改进。按照 XLM 中的掩码语言建模方法(masked language modeling,MLM)用单语数据进行无监督翻译任务。其核心思想是对输入的句子进行掩码和随机词替换的加噪操作,这一点与 BERT[14]类似。然后训练模型提取句子的特征从而在解码端还原输出原始的句子。此外,本文借鉴了文本匹配和情感分析方法,采用交叉注意力机制对两个输入进行特征融合,将模型的编码器结构分为两个部分,把二者的输出特征信息进行融合。本文所提出的模型基于传统无监督神经机器翻译架构,主要对编码器部分进行改进,在输入端进行加噪和共享词典的构造,即图 3 中的 Vocab,编码器部分采用伪孪生网络的框架,其中 Transformer′是基于 XLM 底层的 Transformer 改进后的网络,在 Attention 层和 FFN 层之间引入

了一种噪声过滤门机制,并且编码器中的两个网络用 XLM-encoder 的参数进行共享,解码器之间用 XLM-decoder 的参数进行共享,其余部分为反向翻译训练的过程,训练目标为最小化重构序列与原始序列的误差,模型整体架构如图 3 所示。其中 E_{tgt}、D_{tgt} 分别表示目标语言的编码器和解码器;E_{src}、D_{src} 分别表示源语言的编码器和解码器。

2.1 文本预处理

定义源语言序列为 $\boldsymbol{X}_{src}=[x_{s1},x_{s2},\cdots,x_{si},\cdots,x_{sn}]$,$n$ 为源语言序列长度;目标语言序列为 $\boldsymbol{X}_{tgt}=[x_{t1},x_{t2},\cdots,x_{ti},\cdots,x_{tm}]$,$m$ 为目标语言序列长度。输入端先进行共享词典的构建,采用 Sennrich 等[20]提出的 BPE(byte pair encoding)方法对 $\boldsymbol{X}_{src}$ 和 $\boldsymbol{X}_{tgt}$ 的数据进行混合处理。首先,在数据集上采样,并且保持采样的随机概率不变;其次,将采样的结果进行融合,形成共享字词单元;再次,对输入序列进行加噪,此处采用的加噪方式是以概率 P 随机丢弃输入句子中的单词;最后,做 embedding 词嵌入处理,将处理后的源语言和目标语言的特征向量分别记为 $\boldsymbol{e}_{src}$ 和 $\boldsymbol{e}_{tgt}$,作为输入结果传输到编码器中。

2.2 伪 Transformer 编码

模型的编码器分为了两个部分,一部分是 XLM 模型中的 Transformer 原生网络,另一部分是改进的 Transformer′伪孪生网络,二者之间参数共享。在 Transformer′中使用两个通道来处理注意力输出,其中一条通道经过 Sigmoid 函数[21]处理得到去噪门控信号,将该信号与另一条通道的输出做点乘得到注意力输出,最后经过前馈神经网络层得到 Transformer′部分的最终输出。以源语言序列 $\boldsymbol{X}_{src}$ 为例(目标语言序列 $\boldsymbol{X}_{tgt}$ 与之计算方式一致,这里仅列举 $\boldsymbol{X}_{src}$),具体计算过程如下:

$$\boldsymbol{O}_{src}=\text{Attention}(\boldsymbol{e}_{src},\boldsymbol{e}_{src},\boldsymbol{e}_{src}); \tag{1}$$

$$\boldsymbol{O}^1_{src}=W_1\boldsymbol{O}_{src}+b_1; \tag{2}$$

$$\boldsymbol{O}^2_{src}=W_2\boldsymbol{O}_{src}+b_2; \tag{3}$$

$$\boldsymbol{O}^3_{src}=\text{Sigmoid}(\boldsymbol{O}^1_{src}); \tag{4}$$

$$\boldsymbol{P}_{src}=\boldsymbol{O}^2_{src}\cdot\boldsymbol{O}^3_{src}; \tag{5}$$

$$\boldsymbol{Z}'_{src}=\text{FNN}(\boldsymbol{P}_{src})。 \tag{6}$$

式中:W_1、W_2、b_1、b_2 均为可学习的参数;$\boldsymbol{O}_{src}$ 为注意力输出;$\boldsymbol{O}^1_{src}$ 为第 1 个通道经线性层得到的特征向量;$\boldsymbol{O}^2_{src}$ 为第 2 个通道经线性层得到的特征向量;$\boldsymbol{O}^3_{src}$ 为经噪声过滤所得的去噪门控信号;$\boldsymbol{P}_{src}$ 为两条通道相乘得到的注意力输出;$\boldsymbol{Z}'_{src}$ 为 $\boldsymbol{P}_{src}$ 经前馈神经网络层处理得到的最终输出。将 Transformer 原生网络编码结果的输出记为 $\boldsymbol{Z}_{src}$,则编码器整体输出如下:

$$\boldsymbol{d}_{src}=\boldsymbol{Z}'_{src}+\boldsymbol{Z}_{src}。 \tag{7}$$

式中:$\boldsymbol{d}_{src}$ 表示从编码器部分向解码器传入的最终输出。此外,为使得编码器输出特征更接近真实值,采用最小化均方差来优化 Transformer 和 Transformer′的融合特征。

2.3 解码和优化

解码器部分采用 XLM 的 decoder 结构,此处并

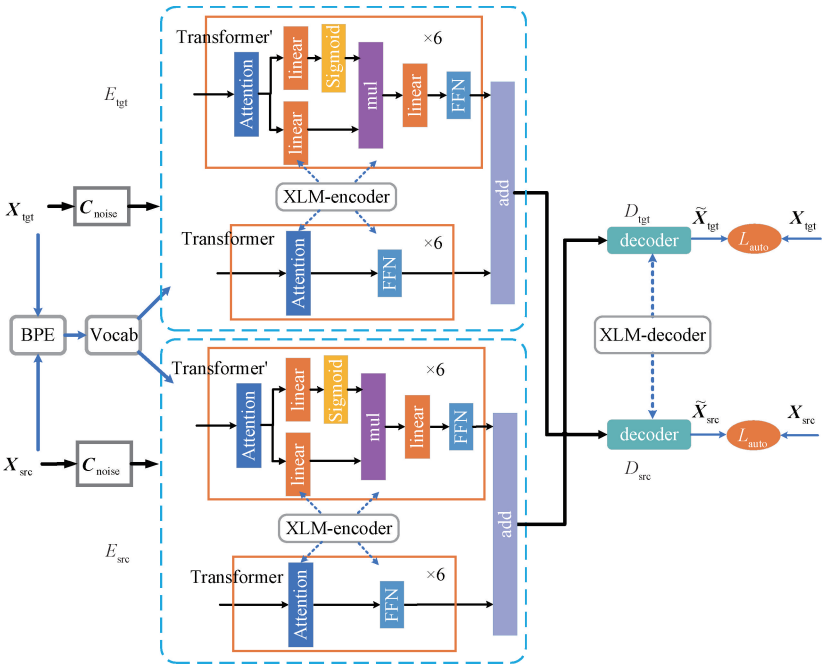

图 3 模型整体架构

Figure 3 Overall architecture of the model

未做出改进,其中源语言和目标语言的解码器之间通过 XLM-decoder 参数共享。decoder 输出结果为重构的译文 $\widetilde{\boldsymbol{X}}_{\text{src}}$,将原文与重构译文之间的交叉熵[22]作为总损失,逐渐缩小二者之间的误差,采用 Adam 优化器进行优化,源语言和目标语言各自进行。计算过程如下:

$$\widetilde{\boldsymbol{X}}_{\text{src}} = \text{decoder}(\boldsymbol{d}_{\text{src}}); \tag{8}$$

$$\boldsymbol{L}_{\text{auto}} = -\sum_i p(x_i)\log q(\tilde{x}_i)。 \tag{9}$$

式中:$\widetilde{\boldsymbol{X}}_{\text{src}}$ 表示从解码端生成的重构译文;L_{auto} 表示反向翻译损失;$p(x_i)$ 为源语言序列 $\boldsymbol{X}_{\text{src}}$ 经归一化得到的概率分布;$q(\tilde{x}_i)$ 为重构译文 $\widetilde{\boldsymbol{X}}_{\text{src}}$ 经归一化得到的概率分布。$\boldsymbol{L}_{\text{auto}}$ 的结果优化后再更新模型参数,使得损失达到最小。

3　实验与分析

3.1　数据集构成

本文采用国际翻译探讨会(workshopon machine translation,WMT)中的 WMT 16 英语-德语(En-De)数据集、WMT 14 英语-法语(En-Fr)数据集、WMT 16 英语-罗马尼亚语(En-Ro)数据集作为训练数据集。英语-德语实验分别使用 En-De newsdev2016 和 En-De newstest2016 作为验证集和测试集;英语-法语实验分别使用 En-Fr newsdev2014 和 En-Fr newstest2014 作为验证集和测试集;英语-罗马尼亚语实验分别使用 En-Ro newsdev2016 和 En-3Ro newstest2016 作为验证集和测试集。实验中训练所用的英语单语规模为 1.79×10^6 条数据,其他 3 种语言单语规模为 6.5×10^5 条数据。

3.2　实验环境和参数设置

本文的实验所采用的编程语言为 Python3.8,使用 NVIDIA GeForce RTX4090 显卡,显存大小为 24 GB,深度学习框架为 Pytorch 1.8.1,CUDA 11.1。实验中编码器结构的两个部分均包含 6 个编码层,使用 Adam 优化器,β 系数分别选用 0.90,0.98,实验训练过程中的相关参数如表 1 所示。

表 1　实验训练参数

Table 1　Experimental training parameter

参数	取值	参数	取值
批处理大小	32	隐藏层维度	1 024
训练轮数	300	Transformer 层数	6
学习率	0.000 1	多头注意力头数	8
Dropout	0.3	最大句长	64

3.3　评估指标

本文采用 *BLEU*[23]作为评价模型性能的指标。其计算公式如下:

$$BLEU = BP\cdot\exp\left(\sum_{n=1}^{N} w_n\log P_n\right)\times 100\%。 \tag{10}$$

$$BP = \begin{cases} 1, m > r; \\ e^{1-r/m}, m \leqslant r。 \end{cases} \tag{11}$$

式中:P_n 为 n-gram 精度;w_n 为对应的权重系数;BP 为句长惩罚系数,是为了防止漏译导致 *BLEU* 值不真实;r 为参考译文长度;m 为模型译文长度。*BLEU* 是一种用于判断机器翻译模型生成的译文和参考译文之间相似程度的评估指标。

3.4　实验结果分析

本文所提出的基于伪孪生网络的无监督神经机器翻译方法(PSUMT)在各语言的翻译任务上均取得良好的效果,相较于基线 XLM 平均提升了 3.5 百分点。不同方法的翻译效果如表 2 所示。

表 2　不同方法的翻译效果

Table 2　Translation effects of different methods

模型	*BLEU*/%					
	德-英	英-德	法-英	英-法	罗-英	英-罗
XLM[7]	34.3	27.0	33.3	33.4	31.8	33.3
MASS[15]	35.2	28.3	34.9	37.9	33.1	35.2
mBART[17]	34.0	29.8	32.1	33.5	30.5	35.0
CBD[24]	36.3	30.1	35.5	38.2	33.8	36.3
SUNMT[25]	34.8	28.1	35.1	37.8	33.9	36.2
QBT[26]	35.5	28.7	35.2	37.8	33.2	35.2
PSUMT	36.5	30.8	35.8	38.4	35.4	37.6

(1) MASS(masked sequence to sequence pre-training)[15]:是基于 Transformer 联合预训练编码器和解码器的预训练模型,采用单语数据进行无监督微调的机器翻译系统。

(2) mBART(masked sequence to sequence pre-training)[17]:是基于多语言的 Sequence to Sequence 去噪自动编码器的方法,在大规模单语语料库上预训练,可以在任何语言对上微调而不需要针对特定的任务做出语言上的修改。

(3) CBD(cross-model back-translated distillation)[24]:是在标准无监督机器翻译中引入跨模型反向翻译蒸馏的方法。通过训练两个双向的无监督机器翻译模型,将二者生成的单语数据进行多次反向翻译用于合成平行语句,进而以有监督方式训练模型,训练数据更为多样化。

(4) SUNMT(self-trainingunsupervised machine translation)[25]:是在基于回译的无监督机器翻译的基础上,引入一种在线自训练的方法,弥补内容和风

格差距,同时使用伪并行数据来模拟推理场景,缩小了模型训练和推理之间的差异。

(5) QBT(quick back-translation)[26]:是一种快速反向翻译方法。将编码器视作生成模型,使用其生成的序列结合原始反向翻译来对解码器进行训练,提高了数据吞吐量和利用率,提升了翻译质量。

以英-德实验为例,记录上述不同机器翻译方法的 *BLEU* 随训练轮次的变化情况,其中训练轮次为 10~50,如图 4 所示。

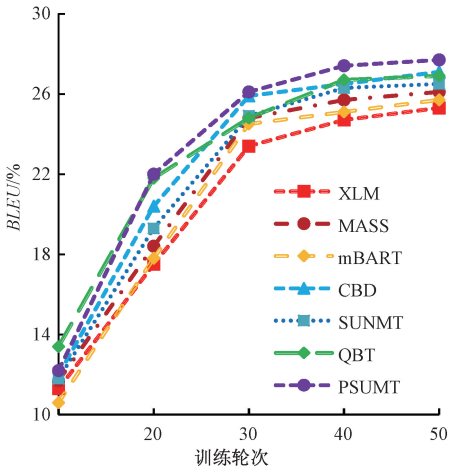

图 4 *BLEU* 随训练轮次变化曲线

Figure 4 *BLEU* with the number of training round

结合表 2 和图 4 可以看出,上述无监督机器翻译算法相较于 XLM 均有一些提升,但本文提出的 PSUMT 方法在 6 个翻译任务上取得的效果明显更好。

从整体上看,语言差异带来的影响不可避免,例如英语-法语和英语-罗马尼亚语获得的 *BLEU* 比法语-英语或者罗马尼亚语-英语高,而在英-德翻译任务中却是相反的表现。

从不同模型的表现情况来看,基线模型 XLM 表现最差,MASS 模型略优于 XLM,其在预训练过程中引入固定长度的句子遮蔽片段,对于一些长难句或复杂句,其捕获的语义信息有限,因此性能表现一般。在训练开销和训练复杂度方面,mBART 模型和 CBD 模型表现欠佳,这是因为 mBART 模型的训练需要大量多种语言数据的支撑,CBD 模型所依赖的两个双向模型在训练时往往会带来更多的噪声信息。相比较之下,SUNMT 方法所采用的在线自训练方式不需要额外的计算来生成伪平行语料,节省了一些资源,但其对翻译目标的质量要求很高。QBT 方法的模型规模相对较小,在训练速度上有较大优势,在小批次训练条件下效果最佳,但能够处理的任务有限,需要与其他模型结合使用,从而使得机器翻译系统的复杂度增加。本文所提模型有效均衡了训练开销和复杂程度,采用类似孪生网络的架构,在一定程度提升了模型的训练速度,同时引入门机制保证了翻译的质量,从而避免噪声带来的一些影响,因此在实验中 PSUMT 模型的效果较好。

3.5 消融实验

为证明本文方法的有效性,从两个方面分别测试机器翻译的效果。一方面,仅用模型中 Transformer′的部分作为编码器,以验证噪声过滤门机制的效果;另一方面,将编码器部分设置为两个完全一样的 Transformer 结构,以验证孪生网络结构的效果。消融实验如表 3 所示。

表 3 消融实验

Table 3 Ablation experiment

模型	*BLEU*/%					
	德-英	英-德	法-英	英-法	罗-英	英-罗
XLM	34.3	27.0	33.3	33.4	31.8	33.3
XLM+门机制	34.6	28.5	33.9	36.0	32.9	36.4
XLM+孪生网络	35.0	29.3	34.6	37.0	32.5	36.0
XLM+PSUMT	36.5	30.8	35.8	38.4	35.4	37.6

由表 3 可知,添加噪声过滤门机制和孪生网络架构的方法相较于 XLM 均有不错的提升,孪生网络架构在德-英、英-德、法-英和英-法翻译任务中的效果比采用噪声过滤门机制的效果更好,但在罗-英和英-罗翻译任务中,却比噪声过滤门机制低了 0.4 百分点,可见在英语和罗马尼亚语之间的无监督翻译任务中,噪声所带来的影响要远高于法语和德语。

为进一步验证模型的泛化能力,本文模拟了实际应用场景中噪声的干扰情况。采用国际口语机器翻译研讨会(international conference on spoken language translation,IWSLT)中的 IWSLT14 德语-英语(De-En)双语数据作为小样本数据集,对测试集中的文本序列以给定概率进行加噪,得到带噪测试数据集。其中,噪声概率 $P = \{0, 0.05, 0.10, 0.20\}$,实验结果如表 4 所示。

表 4 不同噪声概率条件下的测试集性能

Table 4 Test set performance with different noise probability conditions

模型	*BLEU*/%			
	$P=0$	$P=0.05$	$P=0.10$	$P=0.20$
XLM	33.1	29.3	26.5	21.9
XLM+门机制	33.7	31.2	30.1	28.6
XLM +孪生网络	34.3	30.7	28.7	26.2
XLM+PSUMT	35.2	32.3	30.6	29.5

从表 4 可以看出,干扰程度越高,模型性能下降越明显。在不同等级噪声干扰下,基线 XLM 表现最

差并且性能下降最明显,其次为使用孪生网络架构的模型。XLM+门机制有着一定效果,下降速度相对缓慢,PSUMT 效果最优,进一步验证了本文方法在噪声条件下的优越性。

4　结论

本文提出了一种基于伪孪生网络的无监督神经机器翻译方法,该方法沿用了孪生网络的思想,将编码器分为两个参数共享的网络,在结构设计上将其中一个网络引入噪声过滤门机制,有效地将反向翻译迭代过程中产生的噪音信息进行过滤,并提取更为精确的特征信息,以便于解码器输出高质量的译文。在 WMT14 和 WMT16 数据集上的实验结果表明,与其他方法相比,PSUMT 有效提升了机器翻译的质量。

下一步,如何更有效地让编码器-解码器模型学习上下文信息以及如何更有效地提高反向翻译迭代效率将会是无监督机器翻译的重要一环。

参考文献:

[1] SUTSKEVER I, VINYALS O, LE QUOC V. Sequence to sequence learning with neural networks [C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 3104-3112.

[2] MIKOLOV T, KARAFIÁT M, BURGET L, et al. Recurrent neural network based language model [C]//Interspeech 2010. Saint-Malo: ISCA, 2010: 1045-1048.

[3] XU H F, LIU Q H, VAN GENABITH J, et al. Multi-head highly parallelized LSTM decoder 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. Stroudsburg: USAACL, 2021: 273-282.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 5998-6008.

[5] LAMPLE G, CONNEAU A, DENOYER L, et al. Unsupervised machine translation using monolingual corpora only[EB/OL]. (2018-04-13)[2024-10-15]. https://arxiv.org/abs/1711.00043v2.

[6] LI J N, DU Y, DU L. Siamese network representation for active learning[C]//2023 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2023: 131-135.

[7] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining [C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2019: 7057-7067.

[8] KOEHN P, KNOWLES R. Six challenges for neural machine translation[C]//Proceedings of the First Workshop on Neural Machine Translation. Stroudsburg: USAACL, 2017: 28-39.

[9] CONNEAU A, LAMPLE G, RANZATO M, et al. Word translation without parallel data [EB/OL]. (2017-10-11)[2024-10-15]. https://arxiv.org/abs/1710.04087.

[10] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders [C]//Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008: 1096-1103.

[11] 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.
SUN H P, ZHAO T J. A survey on unsupervised neural machine translation[J]. Intelligent Computer and Applications, 2021, 11(2): 1-6.

[12] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: USAACL, 2014: 1532-1543.

[13] MELAMUD O, GOLDBERGER J, DAGAN I. Context2vec: learning generic context embedding with bidirectional LSTM [C]//Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg: USAACL, 2016: 51-61.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2019-05-24)[2024-10-15]. https://arxiv.org/abs/1810.04805v2.

[15] SONG K T, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation [EB/OL]. (2019-06-21)[2024-10-15]. https://arxiv.org/abs/1905.02450v5.

[16] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: USAACL, 2020: 8440-8451.

[17] LIU Y H, GU J T, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[18] MA S M, DONG L, HUANG S H, et al. DeltaLM: en-

coder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders [EB/OL]. (2021-08-18)[2024-10-15]. https://arxiv.org/abs/2106.13736v2.

[19] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a siamese time delay neural network[C]// Advances in Neural Information Processing Systems 6: Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 1993: 737-744.

[20] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: USAACL, 2016: 1715-1725.

[21] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. New York: ACM, 2017: 1243-1252.

[22] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. North American Chapter of the Association for Computational Linguistics, 2018, 39(10):1508-1520.

[23] PAPINENI K, ROUKOS S, WARD T, et al. BLEU[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 311-318.

[24] NGUYEN X P, JOTY S R, KUI W, et al. Cross-model back-translated distillation for unsupervised machine translation[EB/OL]. (2021-05-24)[2024-10-15]. https://arxiv.org/abs/2106.13736v2.

[25] HE Z W, WANG X, WANG R, et al. Bridging the data gap between training and inference for unsupervised neural machine translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: USAACL, 2022: 6611-6623.

[26] BRIMACOMBE B, ZHOU J W. Quick back-translation for unsupervised machine translation[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg: USAACL, 2023: 8521-8534.

Unsupervised Learning Multilingual Neural Machine Translation Based on Pseudo-siamese Network

DU Liming[1], QU Dan[1,2], ZHANG Chuancai[2], XI Yangli[1]

(1. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China; 2. Laboratory for Advanced Computing and Intelligence Engineering, Information Engineering University, Zhengzhou 450001, China)

Abstract: When unsupervised neural machine translation was trained with monolingual data, it inevitably brought a lot of noise information. The errors of the machine translation model accumulated continuously during the training iteration process, affecting the translation effect. To solve this problem, in this study an unsupervised neural machine translation method was proposed based on pseudo-siamese network on the basis of cross-lingual pre-training model (XLM). The model encoder was divided into two modules, in which the pseudo-Siamese network part introduced a noise filtering gate mechanism to filter the noise features in the encoding process, so that the model could better learn the mapping relationship between the source language and the target language. The experimental results showed that in the interactive translation task between English, German, French, and Romanian, the proposed method had an average improvement of 3.5 percentage points compared with the baseline system, which proved its superiority in translation effect. Ablation experiments were used to verify the effectiveness of each component of the model. At the same time, the performance test of the method with different noise conditions was simulated in the German-English translation task, and it also showed good noise resistance.

Keywords: unsupervised machine translation; pseudo-siamese network; monolingual data; noise filtering gate mechanism; cross-language pretraining model