文章编号：1003-0077(2022)09-0057-10

抑制无监督神经机器翻译模型退化的简单方法

吴 霖[1,2]，陈杭英[1,2]，李 亚[1]，余正涛[1,2]，杨晓霞[1,2]，王振晗[1,2]

（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650500；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650500）

摘 要：在中英语料下复现 Facebook 提出的无监督神经机器翻译方法时，我们发现模型出现了退化现象。该文分析了退化的可能原因并提出三种简单方法来抑制模型退化。方法一，遮蔽非目标语输出；方法二，双语词典逐词翻译退化译文；方法三，在训练过程中，添加 10 万句对的平行语料。结果显示，三种方法都能有效抑制模型退化。在无监督条件下，方法二的性能更好，BLEU 值为 7.87；在 10 万语料的低资源条件下，方法一效果更好，BLEU 值为 14.28，该文还分析了产生此现象的原因。

关键词：无监督神经机器翻译；低资源；模型退化

中图分类号：TP391 **文献标识码**：A

Simple Methods to Prevent Model Degeneration in Unsupervised Neural Machine Translation

WU Lin[1,2], CHEN Hangying[1,2], LI Ya[1], YU Zhengtao[1,2], YANG Xiaoxia[1,2], WANG Zhenhan[1,2]

(1.Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China; 2.Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: Model degeneration appears in reproducing the unsupervised neural machine translation method proposed by Facebook AI research in the context of Chinese and English language pairs. We propose three simple yet effective methods to solve this problem. First, we mask the non-target words in the translation. Second, we use a dictionary to translate the degenerated machine translations into the target language. Third, we add some parallel corpus (100k parallel sentences) into the training process. Experimental results show that all three methods can effectively prevent the model from degeneration. In total unsupervised setting, the 2nd method has a better BLEU score of 7.87. With 100k parallel sentences, the 1st method is better with a BLEU score of 14.28.

Keywords: unsupervised neural machine translation; low resource; model degeneration

0 引言

随着神经机器翻译（Neural Machine Translation，NMT）研究的不断深入，越来越多的学者开始关注平行语料不足时如何实现高质量的翻译。针对这一研究热点，Facebook 的 Lample 等[1-3]从 2018 年初至 2019 年初连续发表了三篇仅使用单语语料来实现 NMT 的论文。短短 1 年内，在英法数据集上其 BLEU 值从 15.05 提高到了 33.40[1-3]。该实验结果表明，用无监督的方法几乎达到了监督学习的性能。因此，在英法和其它语对上复现其工作是非常值得的。

在英法数据集上，我们能够复现论文中 33.40 BLEU 值。然而，在中英数据集上尝试了多种训练方案后，BLEU 值都不超过 2，而且随着训练迭代次数的增加，BLEU 值不升反降。在仔细分析训练数据后，我们发现模型出现了退化现象。在中英（英中）翻译时，

收稿日期：2021-01-20 定稿日期：2021-11-05

基金项目：国家自然科学基金（61732005，61761026，61672271）；国家重点研发计划（2019QY1801）；云南省重大科技专项计划项目（202002AD080001）

本该为英文(中文)的译文,实际译文却大都是中文(英文)句子,而且这些句子与输入句子十分相似。

本文分析了模型退化的原因(第2节),并提出了抑制模型退化的三种简单方法(第3节):一是遮蔽非目标语输出;二是逐词翻译退化译文;三是在翻译模型的训练过程中,加入适量的(10万句对)平行语料。

从实验结果来看,这三种方法均能有效抑制模型退化。在中-英测试集上,BLEU值从退化时的0.59提高到4.42(方法一)、7.87(方法二)和12.49(方法三),并且同时使用方法一和方法三的效果最好,BLEU值达到14.28。

1 背景

根据XLM(Cross-lingual Language Model Pretraining)[3]源码①,结合Lample等[1-3]的相关论文,我们总结了无监督神经机器翻译(Unsupervised Neural Machine Translation, UNMT)的主要原理,以方便读者理解该模型在中英语对上退化的原因。

1.1 模型结构

UNMT模型[1-3]由用Transformer[4]实现的编码器E与解码器D组成。源语言与目标语言共用同一个编码器与解码器。编码器负责将源语言s和目标语言t的句子映射到同一隐含向量空间。$E_s(E_t)$表示编码器输入为源(目标)语言。解码器负责将隐含向量空间中的句子解码到对应语言。由于共用解码器,解码器的第一个输入指定输出语种。$D_s(D_t)$表示解码到源(目标)语言。为了叙述方便,我们用$P_{s\to s}=E_s\cdot D_s$、$P_{t\to t}=E_t\cdot D_t$、$P_{s\to t}=E_s\cdot D_t$和$P_{t\to s}=E_t\cdot D_s$分别表示源到源、目标到目标、源到目标和目标到源语言的翻译。

1.2 预训练模型和翻译模型的初始化

借助于BERT[5]的成功,XLM[3]在大量单语语料上用MLM(Masked Language Model)任务来初始化整个编码器和解码器,而不是仅仅初始化词嵌入层[1-2]。结果显示,预训练能够大幅提高UNMT的性能[3]。

1.3 自编码去噪

自编码去噪(Denoising Autoencoding)的训练目标是让模型学习如何从被噪声污染的句子中恢复出原始句子,其目标函数如式(1)所示。

$$L^{\text{lm}}=\mathbb{E}_{x\in S}[-\log P_{s\to s}(x\mid\mathbb{C}(x))]+\mathbb{E}_{y\in T}[-\log P_{t\to t}(y\mid\mathbb{C}(y))] \quad (1)$$

其中,$\mathbb{C}$是一个噪声模型,由一组随机变换操作实现。变换包括随机丢弃一些词和随机交换相邻词的位置。XLM用该任务训练$P_{s\to s}$和$P_{t\to t}$语言模型。

1.4 回译

回译[6](Back Translation)力图把困难的无监督学习转化为更容易的伪监督学习[2]。我们使用$u*(y)$表示把目标语言句子$y\in T$回译为源语言的结果,即$u^*(y)=\text{argmax}P_{t\to s}(u\mid y)$;用$v^*(x)$表示把源语言句子$x\in S$回译为目标语言的结果,即$v^*(x)=\text{argmax}P_{s\to t}(v\mid x)$。这样就可以用伪平行句对$(u^*(y),y)$、$(v^*(x),x)$来训练两个语言方向上的模型,其目标函数如式(2)所示。

$$L^{\text{back}}=\mathbb{E}_{y\in T}[-\log P_{s\to t}(y\mid u^*(y))]+\mathbb{E}_{x\in S}[-\log P_{t\to s}(x\mid v^*(x))] \quad (2)$$

XLM用该任务训练$P_{s\to t}$和$P_{t\to s}$翻译模型。

式(1)中的L^{lm}与式(2)中的L^{back}的加权和构成了整个UNMT模型的目标函数。

1.5 词的表示

在Lample等[1]第一次提出无监督神经机器翻译时,虽然其构架共享编码器,但它使用分离的词表,即源语言与目标语言有各自独立的词表。这不利于把源语言与目标语言映射到同一隐含空间。因此,在其后续工作[2-3]中,源语言与目标语言共用同一词表(joint tokenization)。此外,为了防止低频词损害翻译性能,Lample等[2-3]用BPE[7]把词表中的单词分解为出现频率更高的子词(subword)。子词是XLM网络的最小输入单元。

在用BPE处理共享词表后,可能存在子词同时出现在源语言与目标语言中的现象,我们称这些子词为共享子词。在消融实验中,我们把一部分英法共享子词根据输入语种映射到各自独立的词向量中,称这些子词为分离的共享子词。

2 模型退化机理的分析

2.1 英法语对与中英语对的主要差别

由于英语与法语同属印欧语系,而英语与中文

① https://github.com/facebookresearch/XLM

属于不同的语系，因此，与中英语对相比，英法语对中的两种语言要更类似一些，其语法结构要更接近一些。此外，两组语对的另外一个主要差别是，英法语对存在数量比较大的共享 BPE 子词，而中英语对基本不存在共享 BPE 子词①。

为了测试共享子词对 XLM 模型的影响，我们设计了本消融实验。在英法语对上，我们把一定比例的共享子词分离表示为英语子词和法语子词。这样对于翻译模型而言，这些分离的子词就是没有关联关系的独立子词。我们观察共享子词的分离比例与翻译模型的 BLEU 值间的关系，结果如图 1 所示。

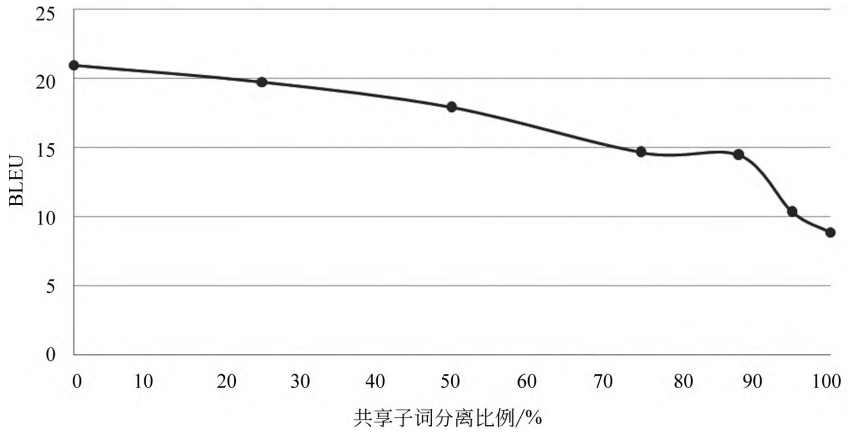

图 1　英法语对上子词分离比例对 BLEU 值的影响

2.2　退化机理的分析

要让 Lample 等[3]提出的 UNMT 模型成功运转，必须实现两个关键步骤：第一，编码器能够把输入的两种语言映射到同一隐含空间，而不是隐含空间中的两个分离子空间；第二，解码器要能够根据输入的语种标签，把隐含空间中的句子解码为对应的目标语种。两者缺一不可。

根据目前的实验结果，在英法和中英语对上，预训练都可以在编码器中把两种语言映射到同一隐含空间②，只是映射的质量还有待设计进一步的实验来评估。

而无监督神经机器翻译能够在英法上训练成功，在中英上却发生模型退化的原因出在自编码去噪和回译这两个步骤上。其具体机理，从目前的实验结果来看还不是十分清晰，但我们目前的理解是：

（1）共享子词对降低自编码去噪和回译任务的难度比较重要。从第 2.1 节的结果可以看出，在英法语对上随着共享子词分离比例的提高，BLEU 值从存在全部共享子词时的 20.98，迅速下降到不存在共享子词时的 8.85。这说明共享子词的确对 XLM 的成功运转很重要。但是当我们分离了所有共享子词后，在英法语对上模型并没有出现退化现象，而仅仅是 BLEU 值下降到 8.85。这说明即使没有共享子词，由于英法是同源语言，其翻译难度与非同源的中英语对相比还是要容易一些。而中英语对，在我们的实验中，如果不使用抑制模型退化的三种方法，中英语对的 BLEU 值从没有超过 2，存在比较严重的模型退化现象。中英语对没有共享子词，所以无法做分离共享子词的消融实验。但是，第 3.2 节中英 UNMT＋C 的结果，其实质是在语义相近的子词间建立联系。当这个联系部分建立起来时，其 BLEU 值提升到 6－7 左右③。另外，对照中英 UNMT＋M 的 BLEU 值 4.42，说明 XLM 模型在没有额外对齐信息的帮助下，仅仅依靠单语语料，还是能够粗略地对齐源语言与目标语言，虽然对齐的效果还不够理想，并且对齐效果还需要遮蔽方法来激活，否则模型就会退化。因此，结合第 2.1 节英法上的消融实验和第 3.1 节中英上的 UNMT＋M 与第 3.2 节中英上的 UNMT＋C 实验来看，共享子词对无监督神经机器翻译的性能还是比较重要的，因为共享子词在源语言与目标语言间建立了天然不变的对齐关系。

① 中英语料中的确存在少量的共享子词，其主要来源是阿拉伯数字和中文语料中出现的一些英文单词。

② 感谢匿名审稿人推荐的两篇论文[8-9]，让我们对实验结果做出了更正确的解读。我们判断预训练能够在中英语对上把两种语言映射到同一隐含空间的依据是这两篇论文[8-9]的共同结论和以下结果：我们查看了停止预训练后，在进行少量自编码去噪和回译训练后的中英译文和英中译文，发现译文大致都是中英文各半。如果结合中文和英文的意思来看，这个混合的译文保留了输入语料的语义，但是单独看中文或者单独看英文，语义是不完整且互补的。

(2) 语言对是否相近,可能也是比较重要的因素。在做消融实验前,我们曾经预期当英法语对上不存在共享子词时,BLEU应该很低,实验结果却是意外的8.85。对此,我们的理解是英法语对由于同源,语法结构比较类似,因此其完成自编码去噪和回译任务的难度较低一些。从而,在不存在共享子词的情况下,其翻译性能还是可接受的。而中英语对,其语法虽然不是完全不一样,但是其差异相较英法语对要大一些,因此其完成自编码去噪和回译任务的难度要高一些,从而出现模型退化现象。当然,如果我们能在更多的语对上做实验,画出BLEU值与语言同源度的关系图,我们的这个猜测可能会更有说服力一些。这项工作,将会在下一步工作中推进。

总之,无监督神经机器翻译在英法语对上工作、在中英语对上不工作,其问题的根源出现在自编码去噪和回译这两个步骤上。这为解决该问题缩小了排查范围,指明了下一步工作的方向。

3 方法

针对模型出现退化的原因,本文提出了两种解决思路、三种方法来应对模型退化。第一种思路是增加约束,即在回译生成语料阶段,强制系统输出目标语,而不是退化后的源语言。其具体方法有两种:遮蔽非目标语的输出和用双语词典逐词翻译退化译文。第二种思路是创造条件降低用自编码去噪和回译来实现无监督神经机器翻译的难度。其具体方法是在训练语料中加入一定量的平行语料。这是治本的方法,所以实验效果最好,但是引入平行语料违背了UNMT的初衷。

3.1 遮蔽非目标语输出(UNMT+M)①

Transformer[4]解码器的最后一层是softmax,其输出结果是词表里每一个子词的输出概率,传统解码方法输出的是概率最大的子词。当模型退化时,概率最大的子词很有可能不是目标语言的子词,而是源语言的子词。因此,遮蔽方法取出概率最大的前k个子词,按概率值降序排序,逐个判断其语种,把语种为目标语的第一个子词输出。如果这个子词的位置是i,这相当于遮蔽了前$i-1$个源语言子词,因此本方法取名"遮蔽非目标语输出"。

如果前k个子词全部是源语言的,那么我们输出第一个子词,即遮蔽方法不激活。遮蔽实验对比了k为5,10,15和20的结果,其中$k=10$的结果最好,所以本文只报告了$k=10$的结果。当k值过小时,在前k个子词中出现目标子词的概率不高,遮蔽方法大多数时刻处于不激活状态,所以k值过小时效果不好。当k值过大时,遮蔽方法大部分时间处于激活状态,但此时遮蔽方法允许输出概率值非常小的目标子词,这些子词也许跟输入语句毫无关联,因此k值过大时效果也不好。实验结果表明,$k=10$是比较合适的选择。

3.2 逐词翻译退化译文(UNMT+C)②

XLM的回译训练由文件trainer.py中的bt_step函数完成。该函数首先获取一个批次的语料,然后用当前的编码器、解码器回译语料,最后用回译的语料做输入,获取的语料做目标,训练编码器和解码器。

逐词翻译的代码就添加在回译之后,训练之前。此时数据的格式是BPE子词的id号。逐词翻译的目标是把退化译文BPE子词的id号逐词转换为目标语言BPE子词的id号。其具体实现步骤如下:

(1) 用词表把子词id序列转换为待处理BPE子词序列;

(2) 读取一个子词,判断子词的语种,如果是目标语种,则输出该子词,并重复步骤2;如果是源语言语种,则转步骤(3);

(3) 拼接待处理子词并查词典,寻找返回非空子集的最长前缀;

(4) 从返回子集的词中随机抽取一个作为译文词。重复步骤(2),直到处理完所有BPE子词;

(5) 用fastBPE把译文词序列切分为BPE子词。用词表把BPE子词转换为子词id。如果输出子词id序列长度l_o大于输入序列长度l_i,则截断输出序列,只输出前l_i个id。如果输出子词id序列长度l_o小于输入序列长度l_i,则补填充(padding)序列让$l_o=l_i$。

为了加快词典的查询速度,词典存储在一个前缀树中。其叶子节点存储了译文词和译文词在语料中出现的次数。随机抽取时,每个译文词被抽取的概率是该词在语料中的次数除以查询返回的非空子集中所有词的次数之和。所以逐词翻译可以理解为增加约束,强制输出译文为目标语的过程,但同时它也是一个加噪的过程,相同子词在不同批次下的输

① M代表遮蔽的Mask。此方法并非作者原创,它来源于论文投稿时匿名评阅人的评阅意见。

② C代表逐词翻译的Converter。

出结果很有可能是不同的。

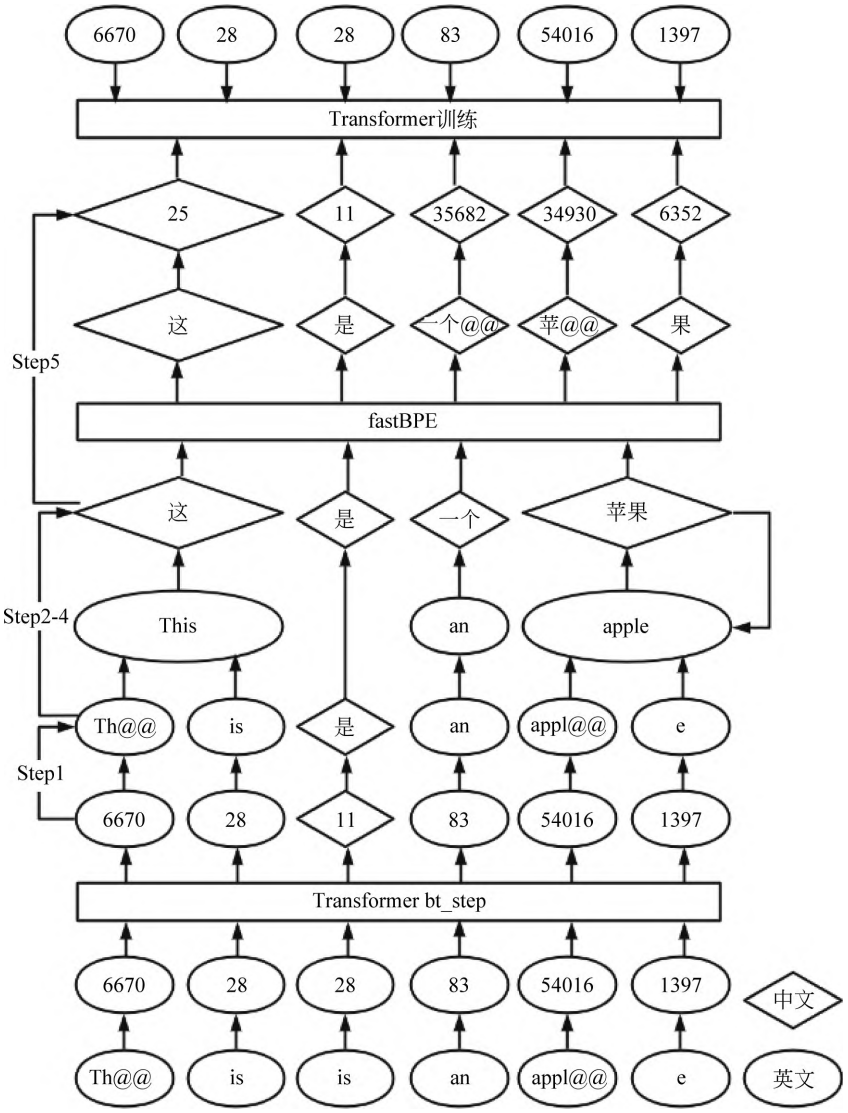

图 2　逐词翻译退化译文的计算过程示例

为了快速完成实验并且减少对实验结果的干扰，我们使用 fast_align[①] 抽取了 100 万句对的平行语料得到双语词典，而没有采用使用单语语料构建双语词典的方法。

3.3　加入少量平行语料（UNMT＋NMT）[②]

在 Facebook 开源的 XLM 实现中，有自编码去噪和回译的训练步骤，两个步骤交替进行。同时 XLM 也提供了使用双语语料的机器翻译训练步骤。代码中一个批次的训练顺序是：先用自编码去噪训练，然后用双语对齐语料进行训练，最后用回译训练。因此，我们不需要改动代码来实现本方案。我们只需要提供正确的运行参数激活双语对齐语料的训练步骤并提供中英语料，即可完成实验。在实验中，我们使用了从 100 万原始平行语料中随机抽取的 10 万个句对。

4　实验

4.1　数据与预处理

英语单语数据使用 WMT14 中 2007 和 2008 年的新闻数据集。汉语单语数据使用厦门大学自然语言处理实验室收集的 2011 年新华社新闻数据集

① https://github.com/clab/fast_align

② 这是方法三在实验结果中的简写名称。UNMT 代表 Unsupervised Neural Machine Translation，NMT 代表使用平行语料的 Neural Machine Translation

(CWMT单语语料)。我们去除了标题,仅留下主体内容。上述语料各抽取500万个句子作为单语训练集。我们从中科院自动化所发布的casia2015数据集(CWMT平行语料)中随机抽取了10万个句对作为中英双语训练集。双语验证集和测试集均来自于WMT18的英-汉双语验证集和测试集。所有数据均使用mosesdecoder[①]提供的tokenizer进行正则化(Normalization)。汉语还使用了斯坦福大学的NLP工具[②]进行了分词。建立BPE词表的操作沿用XLM英法数据集的处理方法和参数。我们用fast_algin抽取casia2015全部句对得到逐词翻译使用的英汉词典。

4.2 训练详情

训练由两步组成:预训练的MLM任务和翻译模型的训练。我们参照XLM论文[3]及其开源代码来选择这两步的超参。由于计算资源有限,并没有尝试其他的超参。

MLM任务的超参如下:6层Transformer模型,8个注意力头,每层1 024个单元,GeLu激活函数。Adam优化器,初始学习率0.0001。Dropout为0.1,注意力dropout为0.1。

翻译模型的超参与MLM任务的超参基本相同,不同的是:自编码去噪的三个噪声控制参数,word_shuffle,word_dropout,word_blank分别设置为3,0.1,0.1;优化器使用带热启动然后按更新数平方根倒数衰减的Adam,beta1与beta2为0.9和0.98,学习率为0.000 1。

在中英语对上,我们使用单块P100 GPU训练MLM任务,大约需要5天时间。训练结束时,两种语言准确率均大于50%,困惑度均小于16。翻译模型大致需要训练2~3天。

4.3 各种训练方法的BLEU值

本文做了九组对比实验:用英汉单语数据训练无监督神经机器翻译模型(简称为UNMT)、无监督遮蔽非目标语输出(UNMT+M)、无监督逐词翻译退化译文(UNMT+C)、无监督遮蔽并逐词翻译(UNMT+MC)、用10万平行语料进行监督训练(NMT)、在无监督翻译模型训练过程中加入10万平行语料(UNMT+NMT)、在无监督翻译模型训练过程中加入10万平行语料还逐词翻译退化译文(NMT+C)、在无监督翻译模型训练过程中加入10万平行语料还遮蔽非目标语输出(NMT+M)、和在无监督翻译模型训练过程中加入10万平行语料还遮蔽非目标语输出并逐词翻译(NMT+MC)。

表1给出了使用MLM初始化编码器与解码器后,九组训练方法在英汉双向翻译任务上验证集与测试集中的在最高第22轮时[③]的BLEU值。

表1 中英语料下九组训练方法在最高第22轮时的BLEU值

实验	验证集		测试集	
	英汉	汉英	英汉	汉英
UNMT	0.21	0.36	0.32	0.59
UNMT+M	3.61	4.17	3.97	4.42
UNMT+C	**6.26**	**7.01**	**7.28**	**7.87**
UNMT+MC	3.76	4.21	3.95	4.1
NMT	10.25	10.16	10.7	10.77
UNMT+NMT	11.46	11.81	11.4	12.49
NMT+C	12.02	12.09	12.31	13.14
NMT+M	12.25	**13.36**	12.49	**14.28**
NMT+MC	**12.7**	13.05	**12.79**	14.25

从表1可以发现,UNMT的BLEU值不足1个点,此时模型发生了退化。当使用UNMT+M(方法一)时,模型退化缓解,BLEU值提升到3.61~4.42。当使用UNMT+C(方法二)时,模型退化进一步缓解,BLEU值提升到6.26~7.87。当使用10万平行句对的监督学习时(NMT),BLEU值提升到10.16~10.77。当在UNMT训练过程中加入10万平行语料(UNMT+NMT,方法三)时,BLEU值进一步提升到11.4~12.49。当在无监督翻译模型训练过程中加入10万平行语料还逐词翻译退化译文(NMT+C,同时使用方法二和方法三)时,BLEU值进一步提升到12.02~13.14。性能最好的是在无监督翻译模型训练过程中加入10万平行语料还遮蔽非目标语输出(并逐词翻译退化译文)(NMT+M同时使用方法一和方法三;NMT+MC同时使用方法一、方法二和方法三),此时BLEU值进一步提升到12.25~14.28。由于NMT+MC比NMT+M的性能提升不明显,而NMT+MC比NMT+M复杂很多,综合方法的简单性和性能,我们认为NMT+M的性能最好。

① https://github.com/moses-smt/mosesdecoder

② https://nlp.stanford.edu/software/stanford-segmenter-4.2.0.zip

③ UNMT、UNMT+M和UNMT+MC,在不到22轮时就退出了。

实验结果表明，单独使用三种方法都能有效抑制模型退化问题。从 BLEU 值上来看，加入 10 万平行语料（UNMT＋NMT）的性能最好，其次是引入双语词典（UNMT＋C），再其次是遮蔽方法（UNMT＋M）。

对比这三个结果，我们的理解是：BLEU 值依赖于输入提供的对齐信息的质量。引入 10 万平行语料，能够建立子词与子词间在不同上下文（context）下的精确关联关系，因此性能最好。引入双语词典，能够建立子词与子词间的与上下文无关的部分关联关系，因此性能居中。遮蔽方法，子词与子词间的依赖关系来源于模型自身，缺乏外力的帮助，因此性能最差。此外，UNMT＋NMT 的性能要优于 NMT，说明基于单语语料的自编码去噪和回译任务的确能够提升 NMT 的性能。

在无监督环境下，遮蔽方法（UNMT＋M）的性能在三种方法中是最差的。意外的是，在加入 10 万平行语料后，遮蔽方法（NMT＋M）超越逐词翻译（NMT＋C）成为性能最好的方法。其原因是，10 万平行语料把源语言子词与目标语言子词间的联系初步建立起来了。如果译文中还出现源语言，NMT＋M 输出概率更低的后续目标语言子词，与概率最大的源语言子词间有关联的可能性是很高的，因为有 10 万平行语料的帮助，并且这个关联是随着上下文不同而会相应变化的，因此它的效果比 NMT＋C 要好。

从实验结果来看，联合使用遮蔽和逐词翻译（＋MC）的性能，相较于逐词翻译（＋C）来说，更接近遮蔽（＋M）方法一些。其原因是，遮蔽方法的代码写在 Transformer 层，逐词翻译的代码写在模型输出外围，因此＋MC 方法是先遮蔽方法产生作用，然后再逐词翻译产生作用。当输出用遮蔽方法变为目标语后，逐词翻译不激活；只有在遮蔽方法无法转化为目标语的输出时，逐词翻译才产生作用。从图 3 来看，（U）NMT＋M 与（U）NMT＋MC 的结果很接近，因此在（U）NMT＋M 中引入逐词翻译对性能没有明显提升。

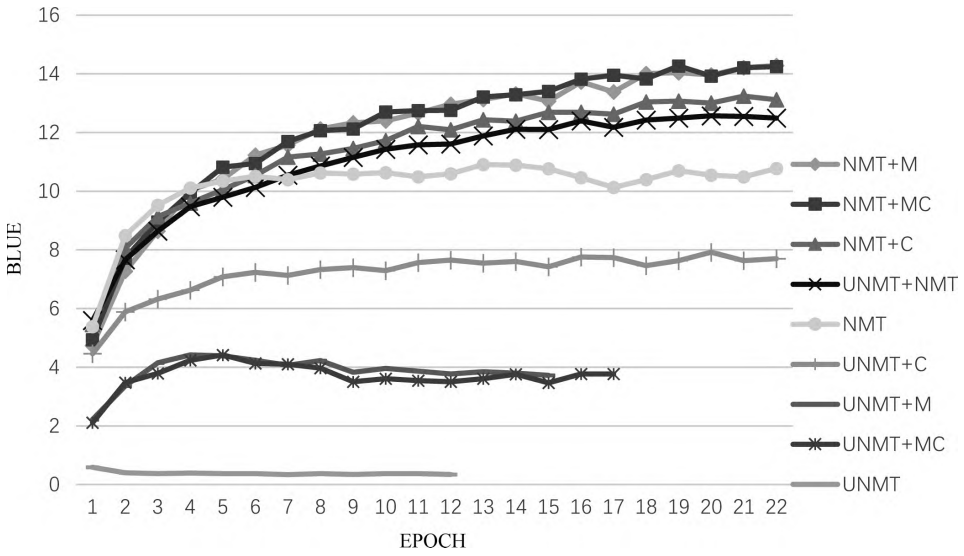

图 3　中-英翻译任务在测试集上的 BLEU 值

图 3 展示了在测试集上，随着训练轮数的增加，汉语到英语 BLEU 值的变化情况。从图中可以看出，遮蔽方法（U）NMT＋M 与遮蔽加逐词翻译（U）NMT＋MC 的性能很接近，两者的曲线交错在一起。如果把（U）NMT＋M 与（U）NMT＋MC 看作同一方法，七种训练方法（UNMT、UNMT＋M、UNMT＋C、NMT、UNMT＋NMT、NMT＋C 和 NMT＋M）BLEU 值的差距是非常明显的，除了训练初期曲线存在交错的情况。因此，BLEU 值的差距能够可靠地反映七种方法间的性能差异。此外，虽然我们在 22 轮时停止了训练，但是此时 UNMT＋C、UNMT＋NMT、NMT＋C、NMT＋M、和 NMT＋MC 还有缓慢的上升趋势。如果我们继续训练，这五种方法的 BLEU 值还会提高。英语到汉语 BLEU 值的变化情况与图 3 基本相同，由于篇幅有限，我们没有展示其结果。

4.4　翻译实例的分析

表 2 展示了九种模型在中英测试集上的翻译实例。从表 2 中可以看出，UNMT 发生了模型退化，

而其他八种方法都能把源语言翻译到目标语种，因此这八种方法都能有效抑制模型退化。

表2 九种翻译模型在测试集上的翻译实例

中文原文	他的家人告诉电视台说，孩子有望康复.	
英文原文	Family tells the station he is expected to recover.	
翻译模型	英译中	中译英
UNMT	Family tells the station he is expected to recover.	他的家人告诉电视台说，孩子有望康复.
UNMT+M	家属告诉记者，他预计今年年底前将退休.	His family told the BBC that the children could be released.
UNMT+MC	老人告诉记者，他预计将恢复正常.	His family told CNN the child could die.
UNMT+C	家庭讲述：派出所有望收回	His family told television that the children are expected to be recovered.
NMT	家庭成员告诉电视台，他的表现很有可能会恢复.	His family told TV that the child was expected to recover from his recovery.
UNMT+NMT	家人告诉车站他有望恢复	His family told television that the child was expected to be recovered.
NMT+C	家庭告诉地铁他有望恢复	His family told television that the child was expected to be recovered.
NMT+MC	家人告诉车站他有望恢复	His family told the TV station the child is expected to be rehabilitation.
NMT+M	家人告诉车站他预计恢复	His family told the television that the child was expected to recover.

然而八种方法的翻译质量存在较大差异。总体来说，使用了平行语料的方法（NMT，UNMT+NMT，NMT+C，NMT+MC，NMT+M）的译文在句子流畅度、句子自然程度和用词的准确性上都要更好一些，这说明平行语料为模型提供了其他方法目前还无法提供的信息，这为无监督方法的研究提供了努力的方向。

基于双语词典（UNMT+C和NMT+C）和基于无监督遮蔽（UNMT+M，UNMT+MC）的方法在用词准确性上都存在问题。困扰双语词典的问题是一词多义。例如，station被翻译为“派出所”（UNMT+C）和“地铁”（NMT+C）而不是“电视台”。这说明忽略上下文信息在多个译文中根据词频随机选择译文是有问题的。无监督遮蔽的问题是：遮蔽原文而选择输出概率更低的词，从原理上就会导致用词不准确。其用词不准的程度比一词多义更加严重。例如，station被翻译为“记者”（UNMT+M，UNMT+MC），recover被翻译为“退休”（UNMT+M），Family被翻译为“老人”（UNMT+MC）等。

而在加入10万平行语料后，遮蔽方法（NMT+MC，NMT+M）的用词准确性有了很大提高。虽然station被误译为“车站”，但要比无监督（UNMT+M，UNMT+MC）译为“记者”更准确一些。此外，其他实体词都被正确翻译了，句子流畅度、自然程度和语义的准确性都有很大的提高。

另外，我们注意到在中译英时，NMT方法存在复述的问题。“recover from his recovery”使用了两个有恢复意思的词，其原因有待进一步的实验分析。

5 相关工作

如何利用单语语料是UNMT的核心问题。Gulcehre等[10]提出浅融合和深融合，利用单语语料训练的语言模型来提升NMT性能。Sennrich等[7]提出回译，利用目标单语语料提升NMT性能。目前，回译已成为低资源NMT的主流技术。

2018年初在ICLR会议上，来自Facebook人工智能实验室的Lample等[1]和西班牙巴斯克大学的Artetxe等[11]同时独立提出了UNMT。2018年中期，Lample等[2]提出用统计机器翻译提升UNMT的性能。2019年，Artetxe等[12]也提出了基于该思路的实现方法。2019年初，Lample等[3]提出用预训练进一步提升UNMT的性能。我们认为，Facebook团队在UNMT领域拥有最好的性能、最新的进展、最深的积累和最快的反应速度，因此本

文在 Lample 等人[3]的基础上开展工作。此外，在结合统计机器翻译和神经机器翻译方面，我们国家的学者也做出了很多很有创意的工作[13-17]，值得借鉴。

模型构架的改进也是值得关注的领域。为了把不同语言映射到同一隐含空间，Lample 等[1-3]和 Artetxe 等[11]都使用了共享编码器的构架。Yang 等[18]提出不同语言使用独立的编码器但共享编码器最后几层权重的模型构架。Zheng 等[19]提出镜像生成神经机器翻译模型（Mirror-Generative NMT），把源语言模型、目标语言模型、源-目标翻译模型、目标-源翻译模型整合到一起联合训练。如何在 Lample 等人[3]的基础上利用最新模型构架的进展也是我们关注的方向之一。

6　总结与展望

本文分析了 XLM[3]在中英语料下出现模型退化的原因，并提出了三种简单的解决方案，分别是遮蔽非目标语输出、逐词翻译退化译文和添加少量的平行语料。实验结果表明，这三种方法均能成功抑制模型退化的问题。对比遮蔽与逐词翻译方法，我们发现在无监督条件下，逐词翻译的性能更好；在有 10 万语料的低资源条件下，遮蔽的性能更好。4.3 节中我们分析了该现象的原因。

在抽取字典的过程中，为了尽快完成实验，我们使用了平行语料。未来我们会探索仅使用单语语料来构建词典的方法，例如，使用跨语言词嵌入[20-23]。

虽然我们成功抑制了模型退化的问题，但是译文的 BLEU 值偏低，未来将研究在低资源条件下进一步提高译文质量的方法。目前，最好结果是 Duan 等[24]发表在 ACL 2020 的工作。我们将研究该方法获得成功的机理，然后探索进一步提升无监督神经机器翻译的有效途径。

参考文献

[1] LAMPLE G, CONNEAU A, DENOYER L, et al. Unsupervised machine translation using monolingual corpora only [C]//Proceedings of the International Conference on Learning Representations, 2018: 1-12.

[2] LAMPLE G, OTT M, CONNEAU A, et al. Phrase-based and neural unsupervised machine translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 5039 - 5049.

[3] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining [C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019.

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

[5] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

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

[7] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015: 1715-1725.

[8] WU S, CONNEAU A, LI H, et al. Emerging cross-lingual structure in pretrained language models [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6022-6034.

[9] KARTHIKEYAN K, WANG Z, MAYHEW S, et al. Cross-lingual ability of multilingual BERT: An empirical study [C]//Proceedings of the International Conference on Learning Representations, 2020.

[10] GULCEHRE C, FIRAT O, XU K, et al. On integrating a language model into neural machine translation [J]. Computer Speech and Language, 2017, 45(1): 137-148.

[11] ARTETXE M, LABAKA G, AGIRRE E. Unsupervised neural machine translation [C]//Proceedings of the 56th International Conference on Learning Representations, 2018: 46-55.

[12] ARTETXE M , LABAKA G , AGIRRE E. An effective approach to unsupervised machine translation [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 194-203.

[13] REN S, ZHANG Z, LIU S, et al. Unsupervised neural machine translation with SMT as posterior regularization [C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019: 241-248.

[14] WANG X, LU Z, TU Z, et al. Neural machine

translation advised by statistical machine translation [C]//Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, 2017: 1209-1213.

[15] HE W, HE Z, WU H, et al. Improved neural machine translation with SMT features [C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 151-157.

[16] WANG X, TU Z, XIONG D, et al. Translating phrases in neural machine translation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1421-1431.

[17] WANG X, TU Z, ZHANG M. Incorporating statistical machine translation word knowledge into neural-machine translation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(12): 2255-2266.

[18] YANG Z, CHEN W, WANG F, et al. Unsupervised neural machine translation with weight sharing [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 46-55.

[19] ZHENG Z X, ZHOU H, HUANG S J, et al. Mirror-generative neural machine translation [C]//Proceedings of 8th International Conference on Learning Representations, 2020.

[20] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 789-798.

[21] LI Y, LUO Y, LIN Y, et al. A simple and effective approach to robust unsupervised bilingual dictionary induction [C]//Proceedings of the 28th International Conference on Computational Linguistics, COLING, 2020: 5990-6001.

[22] **张檬,刘洋,孙茂松.基于非平行语料的双语词典构建** [J].**中国科学:信息科学**, 2018, 48(5): 564-573.

[23] DOU Z, ZHOU Z, HUANG S. Unsupervised bilingual lexicon induction via latent variable models [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 621-626.

[24] DUAN X, JI B, JIA H, et al. Bilingual dictionary based neural machine translation without using parallel sentences [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 1570-1579.

吴霖(1976—),**博士,讲师,主要研究领域为机器翻译、自然语言处理、机器学习。**
E-mail: 52038994@qq.com

李亚(1978—),**通信作者,博士,副教授,主要研究领域为自然语言处理。**
E-mail: 59515091@qq.com

陈杭英(1998—),**硕士研究生,主要研究领域为机器翻译。**
E-mail: 1940685132@qq.com