基于去噪的无监督神经机器翻译比基于 MASS 的无监督神经机器翻译更能适应语序差异

Tamali Banerjee孟买
印度理工学院，印度
tamali@cse.iitb.ac.in

Rudra Murthy V
IBM 研究实验室，印度
rmurthyv@in.ibm.com

Pushpak Bhattacharyya
孟买印度理工学院，印度 pb@cse.iitb.ac.in

arXiv:2303.01191v1 [cs.CL] 2 Mar 2023

摘要

我们旨在探究采用自监督预训练的无监督神经机器翻译方法是否能够适应不同语言对之间的语序差异。为此，我们比较了使用相同自监督预训练目标进行预训练的两个模型。第一个模型在语序不同的语言对上训练，第二个模型则在相同语言对上训练，但将源语言重新排序，使其语序与目标语言一致。理想情况下，能够适应语序差异的无监督神经机器翻译方法在这两种设置下的性能应当没有明显差异。本文研究了两种基于自监督预训练的无监督神经机器翻译方法，即掩码序列到序列预训练（MASS）（不含乱序噪声）和去噪自编码器（DAE）（含乱序噪声）。

我们针对五种英语→印度语言对开展实验（即 en-hi、en-bn、en-gu、en-kn 和 en-ta），其中源语言的语序为主谓宾（SVO），目标语言的语序为主宾谓（SOV）。我们发现，对于这些语言对，基于DAE的无监督神经机器翻译方法在翻译准确率方面始终优于MASS。此外，通过重新排序来弥合词序差异，可以提高基于 MASS 的无监督神经机器翻译模型的翻译准确率，但无法提高基于DAE的无监督神经机器翻译模型的翻译准确率。这一发现表明，基于DAE的无监督神经机器翻译比基于 MASS 的无监督神经机器翻译更能适应语序差异。DAE方法中的词语打乱噪声可能是其能够适应语序差异的原因。

1 引言

近年来，基于自监督预训练的无监督神经机器翻译方法（无监督神经机器翻译（Conneau 和 Lample，2019年；Song 等人，2019年；Liu 等人，2020年）已成为主流。这些无监督神经机器翻译模型分两个阶段训练，即（i）自监督预训练和（ii）迭代回译。自监督预训练阶段通常使用自监督方法，在两种或更多语言的合并单语语料库上训练编码器-解码器 Transformer。第二阶段则采用迭代回译，训练模型实现所有方向的翻译。

这些方法在预训练阶段采用的自监督目标各不相同。例如，Conneau 和 Lample (2019年) 采用掩码语言建模（MLM）目标来训练编码器。在此过程中，输入序列中的部分词元会被掩码，训练目标是预测这些被掩码的词元。MASS（Song 等人，2019年）通过掩码连续词元，并使用解码器生成被掩码的词元。Liu 等人(2020年) 使用去噪自编码器（DAE）对模型进行预训练。所采用的噪声函数包括：（a）移除文本片段并用掩码词元替换；（b）打乱句子顺序。

尽管上述方法在密切相关的语言对上取得了令人鼓舞的结果，但源语言与目标语言之间的差异可能会削弱联合训练带来的益处，导致无监督机器翻译模型的翻译质量不佳。

在自然语言处理（Bhat-tacharyya，2012）的不同阶段，均可观察到两种语言之间的差异。近期，已有多种方法被提出，用于解决语言之间的词汇差异问题（Chronopoulouetal.，2021；Banerjeeet al.，2021；Khatri etal.，2021）。然而，词序差异对无监督神经机器翻译的影响仍未得到探索。此前，Sun 等人（2021）尝试了由 Lample 等人（2018）提出的迭代式无监督神经机器翻译方法

该方法不采用自监督预训练，且被发现对词序差异较为敏感。

最近，Abdou 等人（2022）发现，像 DAE 这样经过打乱的语言模型仍会令人惊讶地体现出关于自然语言词序的信息。

重新排序（与预排序同义）一直被用作《机器翻译》中的一种工具，以弥合不同语言之间的词序差异。虽然Du 和 Way（2017）发现重新排序对于神经机器翻译系统是多余的，Zhao 等人（2018）观察到，重新排序可减少欠翻译情况，从而提升效果。在迁移学习场景中，Murthy V 等人（2019）观察到，在训练父模型之前，将辅助源语言的句子重新排序，使其词序与实际源语言相匹配，可以提升结果。

虽然重新排序工具的应用前景看起来很乐观，但使用这类工具需要满足以下条件之一：（i）需要词序经过重新排列的数据来训练有监督的重新排序工具；或（ii）需要句法分析器和重排序规则来使用基于规则的系统。此外，现有的无监督神经机器翻译方法是否能适应语序差异，尚不明确。

本文旨在探究语序差异对最先进的无监督神经机器翻译系统的影响。为此，我们在语序各不相同的语言对上比较 MASS（Song 等人，2019年）和 DAE（Artetxe等人，2018b）。我们在五种语言对上进行了实验，这些语言对中的源语言语序与目标语言语序并不一致。对于每种语言对，我们比较（i）使用源语言→目标语言数据（原始数据）训练的无监督神经机器翻译模型，与（ii）使用重新排序后的源语言 → 目标数据（重新排序后的数据）训练的无监督神经机器翻译模型。这里，重新排序后的源语言指调整源语言句子的顺序，使其与目标语言的语序一致。前一种情况下存在语序差异，而后一种情况下则消除了语序差异。

实验结果表明，使用原始数据和重排序数据训练的基于去噪自编码器的模型，在翻译准确率方面没有显著差异。这说明，与基于 MASS 的无监督神经机器翻译相比，基于DAE的无监督神经机器翻译能够更好地适应两种语言之间的词序差异。我们还观察到，在大多数语言对中，使用原始数据和去噪自编码器方法训练的无监督神经机器翻译模型，其译文质量优于其他模型。

2 所用方法

我们采用Banerjee 等人（2021）提出的 MASS-static 和 DAE-static 方法来处理词汇差异。对于这两种方法，无监督跨语言嵌入会用于初始化无监督神经机器翻译模型的嵌入层；在整个训练过程中（先预训练，再微调），嵌入层均保持不变（静态）。对于每个词元，模型的输入是跨语言词嵌入与位置编码之和。与原始 MASS 方法不同，这里不添加语言代码。

2.1 MASS目标函数

在MASS目标函数中（Song 等人，2019年），会在输入句子中随机选取一个长度为k（其中k为句子长度的一半）的n元语法词元片段。在该片段中，80%的词元会被掩码，10%的词元会被随机替换为其他词元，剩余10%的词元则保持不变。随后，模型经过训练以生成缺失的n元语法。

2.2 DAE目标函数

这里，我们采用基于DAE的预训练；具体而言，在将输入句子作为输入提供给模型之前，先向其中添加随机噪声，并训练模型生成完整的原始句子。虽然可以采用任意噪声（Liu 等人，2020年），但我们使用Artetxe等人（2018b）提出的词语打乱、词语掩码和词语删除噪声。这里，打乱是指交换相邻词语，从而使句子中的局部顺序发生变化。

2.3 带重新排序的无监督神经机器翻译

为了用重排序数据训练无监督神经机器翻译模型，我们首先对单语数据中的源语言句子进行重排序，使其词序与目标语言一致。随后，使用重排序后的源语言数据和目标语言单语数据训练无监督神经机器翻译模型。在源语言→目标语言方向测试该模型时，我们会先对源语言测试句子进行重排序，再将其输入模型。我们不将该模型用作目标语言→源语言系统，因为它生成的源语言句子是重排序后的形式。

语言	# 训练句子
英语 (en)	54.3 M
印地语 (hi)	63.1 M
孟加拉语 (bn)	39.9 M
古吉拉特语 (gu)	41.1 M
卡纳达语 (kn)	53.3 M
泰米尔语 (ta)	3150 万

语言对	# 有效句子	# 测试句子
英语-印地语	1000	2390
英语-孟加拉语	1000	2390
英语-古吉拉特语	1000	2390
英语-卡纳达语	1000	2390
英语-泰米尔语	1000	2390

表1：噪声移除前的数据集统计信息

语言	# 训练句子
英语 (en)	52.2 M
重排序后的英语 (en)	52.2 M

语言对	# 有效句子	# 测试句子
英语-印地语	711	1781
英语-孟加拉语	711	1781
英语-古吉拉特语	711	1781
英语-卡纳达语	711	1781
英语-泰米尔语	711	1781

表 2：噪声移除后的数据集统计信息

3 实验设置

我们的实验中有四组无监督神经机器翻译模型。其中两组使用 MASS-static 训练，另外两组使用 DAE-static 训练。对于这两种无监督神经机器翻译方法，我们分别训练一个模型使用原始数据，另一个模型使用重排序数据。

3.1 语言和数据集

我们使用了 6 种语言的单语数据，即 英语（en）、印地语（hi）、孟加拉语（bn）、古吉拉特语（gu）、卡纳达语（kn）和泰米尔语（ta）。除英语外，其他语言都属于印度语言，词序为主语-宾语-动词（SOV）。英语属于欧洲语系，词序为主语-动词-宾语（SVO）。在实验中使用的所有语言对中，我们均将英语作为源语言，并将这 5 种印度语言之一作为目标语言，因此共得到 5 个语言对，即 en-hi、en-bn、en-gu、en-kn 和 en-ta。

我们使用 IndicCorp 数据集提供的单语数据（Kakwani 等人，2020年）作为训练数据。我们使用 WAT 2021 共享任务提供的英语-印度语言验证集和测试集（Nakazawa 等人，2021）。数据集的统计信息见表1。

3.2 预处理工具

我们使用*Moses*(Koehn 等人，2007) 对英语语料进行分词，并使用*Indic NLP Library*(Kunchukuttan，2020年) 对印度语言语料进行分词。

在训练、验证和测试数据中，我们使用 CFILT-pre-order 的通用规则（Chatterjee 等人，2014年）对英语句子进行重新排序，使其符合印度语言的词序。

对于 BPE 分词（Sennrich 等人，2016），我们使用 *FastBPE*[1] 在源语言和目标数据上联合进行处理，并将合并操作次数设为 100k。

3.3 跨语言嵌入

我们将经过 BPE 分词的源语言和目标语单语语料合并，使用 *Fasttext* (Bojanowski 等人，2017年) 的 Skip-gram 模型训练嵌入。通过合并词汇表的嵌入，我们得到各语言专属词汇表的嵌入。为了提高两种语言的跨语言嵌入质量，我们采用了 *Vecmap* (Artetxe等人，2018a)，并在实验中使用了该方法。

3.4 重排序噪声去除

重排序工具无法正确重新排序句子时，会产生解析错误。为确保比较公平，我们删除了所有这些单语

[1]https://github.com/glample/fastBPE

评估指标	无监督神经机器翻译方法	不同语言对上的翻译准确率					
		英语-印地语		英语-孟加拉语		英语-古吉拉特语	
		源语言 → 目标语言	目标语言 → 源语言 → 目标语言	源语言	目标语言 → 源语言 → 目标语言	源语言	T → S
BLEU	MASS-static	14.16	14.03	1.51	2.77	5.31	6.25
	MASS-static + R	14.63 (↑ 0.47)	-	3.04 (↑ 1.53)	-	8.62 (↑ 3.31)	-
	DAE-static	**21.03**	**21.89**	2.88	**4.39**	**10.60**	**14.78**
	DAE-static +R	15.22 (↓5.81)	-	**3.27** (↑ 0.39)	-	8.72 (↓ 1.88)	-
CHRF	MASS-static	39.45	46.07	25.39	29.56	31.08	36.87
	MASS-static + R	41.77 (↑ 2.32)	-	28.85 (↑ 3.46)	-	37.45 (↑ 6.37)	-
	DAE-static	**45.63**	**52.21**	28.36	**34.61**	**38.55**	**45.74**
	DAE-static +R	42.18 (↓ 3.45)	-	**29.65**(↑ 1.29)	-	37.58 (↓ 0.97)	-

		英语-卡纳达语		英语-泰米尔语	
		S →T	T → S	S → T	T → S
BLEU	MASS-static	3.08	5.11	1.81	2.70
	MASS-static + R	**4.48**(↑ 1.4)	-	2.57（↑ 0.76）	-
	DAE-static	4.42	**9.40**	2.52	**4.77**
	DAE-static +R	4.18 (↓ 0.24)	-	**2.65** (↑ 0.13)	-
CHRF	MASS-static	30.73	33.83	31.40	28.73
	MASS-static + R	**36.47** (↑ 5.74)	-	34.34（↑ 2.94）	-
	DAE-static	35.89	**40.33**	33.00	**34.48**
	DAE-static +R	36.46（↑ 0.57）	-	**35.52**(↑ 2.52)	-

表 3：针对两个方向的语言对（S 和 T 分别指源语言和目标语言），采用 MASS 预训练和 DAE 预训练的无监督神经机器翻译模型在进行或未进行重新排序（R）时的翻译准确率。粗体数值表示各语言对的最佳得分。括号中的数值表示该模型相较于其上方模型在 BLEU/CHRF 上的提升或下降。

英语（源语言）	一个“新印度”正在崛起。
印地语（参考译文）	一个“新印度”正在崛起。 一个“新印度”正在崛起 |
MASS-static	美国的“新印度”正在增长。 美国的“新印度”正在增长 | “新印度”正在美国发展。
重排序-MASS-static	它的“新印度”将会增长。 它的“新印度”将会增长 | 它的“新印度”即将崛起。
DAE-static	一个“新印度”正在崛起。 一个“新印度”正在崛起 | 一个“新印度”正在崛起。
Reordered-DAE-static	“一个新印度”很可能会增长。 “一个新印度”很可能会增长 | 一个新印度很可能会发展壮大。

图 1：使用不同无监督神经机器翻译模型进行 en→hi 翻译的示例

句子（重排序工具对这些句子产生了解析错误），其余数据见表2。包括原始数据和重排序数据中的句子。对于有效和测试用的平行数据，我们也进行了相同的处理（同时删除了对应的译文）。这使数据规模略有缩减。我们在剩余数据上进行了所有实验。具体细节

3.5 网络与评估

我们使用 MASS代码库（Song 等人，2019年）进行实验。模型训练时的训练轮次规模为 $0.2M$ 步，批量大小为 64 个句子（每批词元数 $3K$）。对于每个预训练和

英语源语言句子	**我们**需要改变这种思维方式。
重新排序后的英语源语言句子	**我们**需要改变这种思维方式。
印地语参考译文	我们 需要改变这种想法。 **我们** 需要改变这种思维方式 |
使用 DAE-static 的翻译	我们 必须改变这种思维方式。 **我们** 必须改变这种思维方式 | 我们需要改变这种思维方式。
使用 Reordered-DAE-static 的翻译	我们不需要改变思维方式。 **我们**不需要改变思维方式 | 我们不需要改变思维方式 [输出中缺少格标记]

图 2：重新排序造成歧义的翻译示例

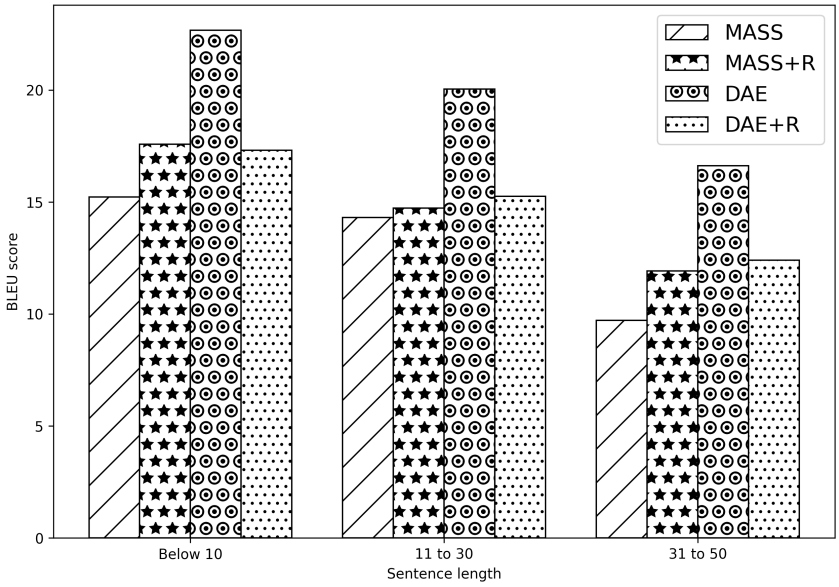

图3：BLEU 分数随句子长度的分布。语言对：en→hi。

微调步骤，我们最多训练模型 50 个训练轮次。不过，如果根据验证集损失判断模型在达到最大训练轮次前已收敛，我们就会停止训练。对于 MASS 预训练，我们使用 0.5 的词掩码比例，即句子中 50% 的词元会被掩码。其他参数均采用代码库中的默认值。对于 DAE 预训练，我们使用词打乱 3、词丢弃 0.1 和词置空 0.1。

我们使用 BLEU（Papineni 等人，2002）和 CHRF（Popovi´c，2016）分数报告翻译系统的性能，并使用 SacreBLEU（Post，2018）进行评估。CHRF（字符 n 元组 F 分数）有助于评估形态丰富的语言，例如印度语言。

4 结果与分析

我们实验中采用或未采用重新排序（R）的无监督神经机器翻译模型的 BLEU 和 CHRF分数如表3所示。我们还报告了加入重新排序后，无监督神经机器翻译模型的翻译质量在 BLEU 和 CHRF分数方面的提升或下降幅度。我们未展示重新排序后的无监督神经机器翻译模型在目标→语言到源语言方向上的结果。这是因为，以目标句为输入时，这些重新排序后的模型会生成其源语言译文，但采用的是重新排序后的形式。对于大多数语言对，使用原始数据训练的 DAE-static 方法无监督神经机器翻译模型取得了最佳结果。图1 中的示例也印证了这一点。DAE-static模型的译文与参考句相似，而其他模型生成的译文则不够恰当。

对于 MASS-static 模型，在重排序数据上训练的模型优于在原始数据上训练的模型。这表明，MASS-static 无监督神经机器翻译无法稳健地应对语

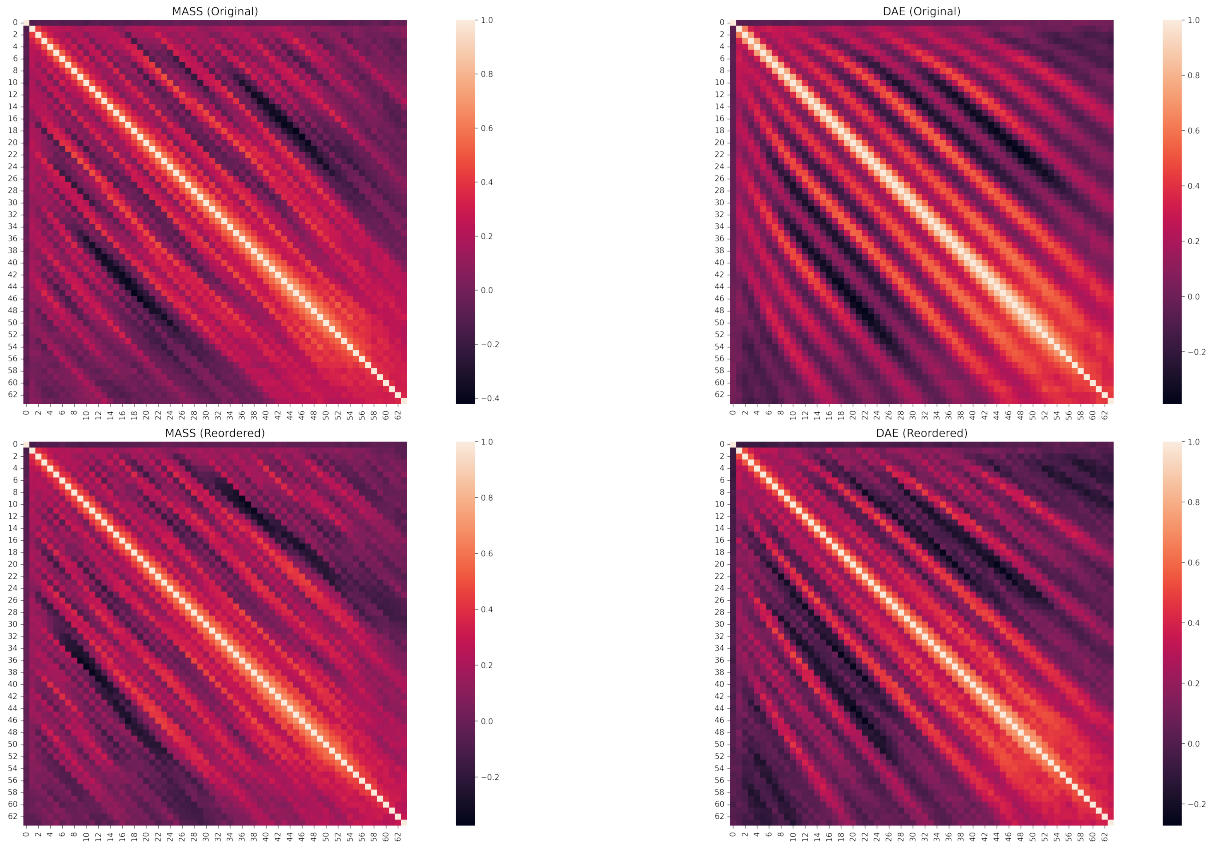

图4：四种无监督神经机器翻译模型（例如 MASS-original、DAE-original、MASS-reordered 和 DAE-reordered）的位置嵌入对比。每张热力图展示了某个无监督神经机器翻译模型的位置嵌入之间的逐位置余弦相似度。图中颜色越浅，表示相似度越高。语言对：en→hi。

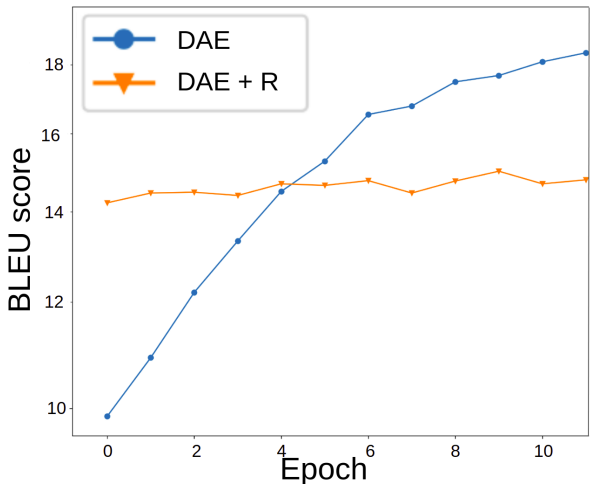

图5：DAE-static 无监督神经机器翻译模型在从第 0 到第 10 个训练轮次进行微调期间，随着训练轮次增加，其翻译 BLEU 分数的变化。对比使用原始数据训练的模型与使用重排序数据训练的模型。语言对：en→hi。

序差异。直观来看，原因在于自监督预训练目标函数

其中不使用打乱。因此，MASS模型会僵化地记忆位置信息。

DAE-static 模型呈现出不同的情况。此处，即使不进行重新排序，模型也能取得最佳结果。这表明 DAE-static 能够适应语序差异。推测 DAE目标函数中的打乱操作，使模型能够灵活地记忆或遗忘位置信息。

令人惊讶的是，对于大多数语言对而言，使用重排序数据训练模型会降低输出的 BLEU 和 CHRF 分数。这是由于在没有格标记的情况下，将主谓宾语序语言重新排序为主宾谓语序所导致的歧义。在图 2 中，我们展示了这样一个例子。在图 2 所示的示例中，主语“we”的参考译文是处于与格的“hameM”。当我们对英语句子进行重新排序时，它仍然是处于主格的“we”，而其常见的印地语译文是“hama”。因此，在重排序模型生成的译文中，我们看到的是“hama”，而不是所需的词

'hameM'。

图 3 显示，MASS 对短句和长句的语序差异更为敏感。DAE-static模型在所有句子长度下的表现都显著优于其他方法。

图 5 显示，当使用重新排序后的数据训练 DAE 时，微调模型仅经过第一个训练轮次就开始生成质量良好的译文。然而，此后模型无法进一步提升；经过若干训练轮次后，使用原始数据训练的模型便超过了它。

图 4 展示了各位置嵌入的逐位置余弦相似度可视化结果。点（i, j）表示第 i^{th}个位置与第 j^{th} 个位置之间的相似度。图中分别展示了 DAE 和 MASS 预训练无监督神经机器翻译模型在经过重新排序和未经过重新排序的数据上的结果。我们观察到，在 MASS 中，位置嵌入与其自身以及局部邻域内相距 2 或 3 个位置的词相似。在 DAE 中，位置嵌入与更大的局部邻域相似，这可能归因于局部打乱噪声。我们还观察到，对于 DAE 模型，在存在语序差异的情况下（即 未经过重新排序），位置嵌入与更多相邻位置相似；而在存在词序相似性的情况下（即 经过重新排序），具有相似位置嵌入的相邻位置数量较少。

5 结论

本文研究了语序差异对基于 MASS 的无监督神经机器翻译和基于DAE的无监督神经机器翻译的影响。为检验这一点，我们考察了在训练前将源数据重新排序，使其词序与目标数据一致，是否能够提高无监督机器翻译模型的翻译准确率。结果表明，基于DAE的无监督神经机器翻译比基于MASS 的无监督神经机器翻译更能适应语序差异。此外，基于DAE的无监督神经机器翻译模型还有一项优势：同一个模型可用于两个方向的翻译（英语到印度语言和印度语言到英语）；而在重排序数据上训练的模型只能用于一个方向的翻译（重排序后的英语 → 印度语言）。

参考文献

Mostafa Abdou、Vinit Ravishankar、Artur Kulmizev 和 Anders Søgaard。2022 年。词序很重要，打乱语言模型对此心知肚明。载于《计算语言学协会第 *60* 届年会论文集（第一卷：长篇论文），第 6907–6919 页，爱尔兰都柏林。计算语言学协会。

Mikel Artetxe、Gorka Labaka 和 Eneko Agirre。2018a。一种用于词嵌入完全无监督跨语言映射的稳健自学习方法。载于计算语言学协会第*56*届年会论文集（第*1*卷：长篇论文），第1卷，第789–798页。

Mikel Artetxe、Gorka Labaka 和 Eneko Agirre。2018b。无监督统计机器翻译。载于*2018*年自然语言处理实证方法会议论文集，第3632–3642页。

Tamali Banerjee、Rudra V Murthy 和 Pushpak Bhattacharyya。2021年。远距离语言的无监督神经机器翻译离不开跨语言嵌入：英语到印度-雅利安语的案例研究。载于第十八届机器翻译峰会：研究分会论文集，第23–34页，线上举行。美洲机器翻译协会。

Pushpak Bhattacharyya。2012。自然语言处理：从计算视角看待歧义、资源限制和多语言环境下的问题。*CSI*计算学杂志，1(2):1–13。

Piotr Bojanowski、Edouard Grave、Armand Joulin 和 Tomas Mikolov。2017年。利用子词信息丰富词向量。计算语言学协会汇刊，5:135–146。

Rajen Chatterjee、Anoop Kunchukuttan 和 Pushpak Bhattacharyya。2014年。基于超标签的机器翻译预排序。载于第*11*届自然语言处理国际会议论文集，第30–38页。

Alexandra Chronopoulou、Dario Stojanovski 和 Alexander Fraser。2021。提升预训练语言模型在无监督神经机器翻译中的词汇能力。载于*2021*年计算语言学协会北美分会会议：人类语言技术，第173–180页，线上。计算语言学协会。

Alexis Conneau 和 Guillaume Lample。2019年。跨语言语言模型预训练。载于神经信息处理系统进展，第7057–7067页。

Jinhua Du 和 Andy Way。2017年。神经机器翻译的预排序：有益还是有害？布拉格数学语言学通报，108(1):171。

Divyanshu Kakwani、Anoop Kunchukuttan、Satish Golla、Gokul N.C.、Avik Bhattacharyya、Mitesh M. Khapra 和Pratyush Kumar。2020年。IndicNLPSuite：面向印度语言的单语语料库、评测基准和预训练多语言语言模型

语言。载于计算语言学协会会议论文集：*EMNLP 2020* 精选论文，第 4948–4961 页，在线。计算语言学协会。

Jyotsana Khatri、Rudra Murthy、Tamali Banerjee 和 Pushpak Bhattacharyya。2021年。简单的词汇差异衔接度量有助于低资源语言的无监督神经机器翻译。《机器翻译》，35(4):711–744。

Philipp Koehn、Hieu Hoang、Alexandra Birch、Chris Callison-Burch、Marcello Federico、Nicola Bertoldi、Brooke Cowan、Wade Shen、Christine Moran、Richard Zens 等。2007年。Moses：用于统计机器翻译的开源工具包。载于第*45*届计算语言学协会年会交互式海报与演示环节论文集，第177–180页。计算语言学协会。

Anoop Kunchukuttan。2020年。印度语言自然语言处理库。https://github.com/anoopkunchukuttan/indic_自然语言处理_library/blob/master/docs/indicnlp.pdf。

Guillaume Lample、Myle Ott、Alexis Conneau、Ludovic Denoyer 和 Marc'Aurelio Ranzato。2018年。基于短语和神经网络的无监督机器翻译。载于*2018*年自然语言处理实证方法会议论文集，第5039–5049页，比利时布鲁塞尔。计算语言学协会。

Yinhan Liu、Jiatao Gu、Naman Goyal、Xian Li、Sergey Edunov、Marjan Ghazvininejad、Mike Lewis 和 Luke Zettlemoyer。2020年。面向神经机器翻译的多语言去噪预训练。计算语言学协会学报，8：726–742。

Rudra Murthy V、Anoop Kunchukuttan、Pushpak Bhattacharyya 等。2019年。解决极低资源语言多语言神经机器翻译中的词序差异问题。载于北美分会年度会议论文集（计算语言学协会）。

Toshiaki Nakazawa、Hideki Nakayama、Chenchen Ding、Raj Dabre、Shohei Higashiyama、Hideya Mino、Isao Goto、Win Pa Pa、Anoop Kunchukuttan、Shantipriya Parida、Ondˇrej Bojar、Chenhui Chu、Akiko Eriguchi、KaoriAbe、Yusuke Oda 和 Sadao Kurohashi。2021 年。 第八届亚洲翻译研讨会概览。载于 《第八届亚洲翻译研讨会论文集（*WAT2021*）》，第 1–45 页，在线。计算语言学协会。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei- Jing Zhu。2002。BLEU：一种用于机器翻译自动评估的方法。载于计算语言学协会第*40*届年会论文集，第311–318页。计算语言学协会。

Maja Popovi´c。2016。深入解析 CHRF：β 参数和 n 元语法权重。载于第一届机器翻译会议论文集：第*2*卷，共享任务论文，第499–504页。

Matt Post。2018 年。 呼吁明确报告 BLEU 分数。载于 《第三届机器翻译会议论文集：研究论文》，第 186–191 页，比利时布鲁塞尔。计算语言学协会。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016 年。使用单语数据改进神经机器翻译模型。载于《计算语言学协会第 *54* 届年会论文集（第 *1* 卷：长篇论文）》，第 1 卷，第 86–96 页。

Kaitao Song、Xu Tan、Tao Qin、Jianfeng Lu 和 Tie- Yan Liu。2019年。MASS：用于语言生成的掩码序列到序列预训练。载于国际机器学习会议，第 5926–5936 页。

Haipeng Sun、Rui Wang、Masao Utiyama、Benjamin Marie、Kehai Chen、Eiichiro Sumita 和 Tiejun Zhao。2021。面向相似和差异较大的语言对的无监督神经机器翻译：一项实证研究。《*ACM* 亚洲及低资源语言信息处理汇刊》（*TAL- LIP*），20(1):1–17。

Yang Zhao、Jiajun Zhang 和 Chengqing Zong。2018。利用预排序进行神经机器翻译。载于第十一届国际语言资源与评测会议（*LREC2018*）论文集，日本宫崎。欧洲语言资源协会（ELRA）。