基于词序扰动的神经机器翻译模型鲁棒性研究

赵彧然，薛傥，刘功申
（上海交通大学网络空间安全学院，上海 200240）

摘　要：预训练语言模型是自然语言处理领域一类十分重要的模型，预训练–微调成为许多下游任务的标准范式。先前的研究表明，将 BERT 等预训练语言模型融合至神经机器翻译模型能改善其性能。但目前仍不清楚这部分性能提升的来源是更强的语义建模能力还是句法建模能力。此外，预训练语言模型的知识是否以及如何影响神经机器翻译模型的鲁棒性仍不得而知。为此，使用探针方法对两类神经翻译模型编码器的句法建模能力进行测试，发现融合预训模型的翻译模型能够更好地建模句子的词序。在此基础上，提出了基于词序扰动的攻击方法，检验神经机器翻译模型的鲁棒性。多个语言对上的测试结果表明，即使受到词序扰动攻击，融合 BERT 的神经机器翻译模型的表现基本上优于传统的神经机器翻译模型，证明预训练模型能够提升翻译模型的鲁棒性。但在英语–德语翻译任务中，融合预训练模型的翻译模型生成的译文质量反而更差，表明英语 BERT 将损害翻译模型的鲁棒性。进一步分析显示，融合英语 BERT 的翻译模型难以应对句子受到词序扰动攻击前后的语义差距，导致模型出现更多错误的复制行为以及低频词翻译错误。因此，预训练并不总能为下游任务带来提高，研究者应该根据任务特性考虑是否使用预训练模型。

关键词：神经机器翻译；预训练模型；鲁棒性；词序
中图分类号：TP393
文献标志码：A
DOI: 10.11959/j.issn.2096−109x.2023078

Research on the robustness of neural machine translation systems in word order perturbation

ZHAO Yuran, XUE Tang, LIU Gongshen
School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: Pre-trained language model is one of the most important models in the natural language processing field, as pre-train-finetune has become the paradigm in various NLP downstream tasks. Previous studies have proved integrating pre-trained language models (e.g., BERT) into neural machine translation (NMT) models can improve translation performance. However, it is still unclear whether these improvements stem from enhanced semantic or

收稿日期：2022−09−27；**修回日期：**2023−03−02
通信作者：刘功申，lgshen@sjtu.edu.cn
基金项目：国家自然科学基金（U21B2020）；上海市科技计划项目（22511104400）
Foundation Items: The National Natural Science Foundation of China (U21B2020), Shanghai Science and Technology Plan (22511104400)

引用格式：赵彧然, 薛傥, 刘功申. 基于词序扰动的神经机器翻译模型鲁棒性研究[J]. 网络与信息安全学报, 2023, 9(5): 138-149.
Citation Format: ZHAO Y R, XUE T, LIU G S. Research on the robustness of neural machine translation systems in word order perturbation[J]. Chinese Journal of Network and Information Security, 2023, 9(5): 138-149.

syntactic modeling capabilities, as well as how pre-trained knowledge impacts the robustness of the models. To address these questions, a systematic study was conducted to examine the syntactic ability of BERT-enhanced NMT models using probing tasks. The study revealed that the enhanced models showed proficiency in modeling word order, highlighting their syntactic modeling capabilities. In addition, an attacking method was proposed to evaluate the robustness of NMT models in handling word order. BERT-enhanced NMT models yielded better translation performance in most of the tasks, indicating that BERT can improve the robustness of NMT models. It was observed that BERT-enhanced NMT model generated poorer translations than vanilla NMT model after attacking in the English-German translation task, which meant that English BERT worsened model robustness in such a scenario. Further analyses revealed that English BERT failed to bridge the semantic gap between the original and perturbed sources, leading to more copying errors and errors in translating low-frequency words. These findings suggest that the benefits of pre-training may not always be consistent in downstream tasks, and careful consideration should be given to its usage.

Keywords: neural machine translation, pre-training model, robustness, word order

0　引言

将预训练语言模型（PLM，pre-trained language model），如 BERT[1]等，融入神经机器翻译模型中是自然语言处理领域十分热门的研究方向[2-5]。融合之后，机器翻译模型能够利用 PLM 学习到丰富语义和句法知识[6-15]，进而提升翻译模型的鲁棒性和翻译效果[16]。但之前的研究更关注改进模型的效果，忽略了对这些提升的来源进行解释，并且缺乏预训练语言模型如何影响翻译模型鲁棒性的考量。

本文首先检验了 BERT 能否改进机器翻译模型的句法能力。使用一些探针对传统的神经机器翻译模型 Transformer[17]（记作 NMT 模型）的编码器和融合了 BERT 的 Transformer 模型（记作 BERT-NMT 模型）的编码器进行测试[10,18]，检验它们分别编码了哪些句法相关的信息。实验结果表明，融合 BERT 后，机器翻译模型编码器在词序建模以及重建依存句法树相关任务上的表现更好。为了验证这些提升确实是由 BERT 带来的，实验过程中使用掩码操作分别遮掩了自注意力模块和 BERT-编码器（BERT-Enc，BERT-encoder）注意力模块。结果证明，BERT 的确能够帮助提高翻译模型对词序进行建模的能力，特别是显著提高模型在双词调换（BShift，bi-gram shift）任务上的表现。

直觉上来讲，这种特性能够帮助机器翻译模型更好地处理源语言句子中的词序扰动攻击。为了检验这一猜想，本文提出了一种以概率 p 交换两个相邻词以改变词序的攻击方法，并使用双语替换评测[19]（BLEU，bilingual evaluation understudy）和 BERTScore[20]对攻击后模型的翻译结果进行评价。在机器翻译工坊（WMT，workshop on machine translation）提供的多个语言翻译任务上进行实验表明，虽然词序扰动的概率不同，BERT-NMT 模型生成的译文质量几乎优于 NMT 模型生成的译文，但在英德翻译任务中，随着扰动概率的增加，使用英文 BERT 反而会给模型生成的译文质量带来消极影响，说明英文 BERT 将破坏翻译模型的鲁棒性。

为了找到使用英文 BERT 训练得到的 BERT-NMT 模型翻译质量下降的原因，比较攻击前后源语言端和目标语言端句子的语义变化。使用 WMT14 英德数据集分别训练得到英德翻译模型和德英翻译模型并进行对比。对比发现，英语句子在受到攻击之后语义变化更为明显，这可能是德语具有比英语更为灵活的词序。而源语言端和目标语言端的语义差别则说明：英语 BERT 难以缩小词序扰动攻击前后源语言端的语义变化。为了细致了解翻译质量下降的具体表现，对比 NMT 模型和 BERT-NMT 模型分别受到攻击之后输出的译文。随着攻击时扰动概率的增加，BERT-NMT 模型倾向于错误地从源语言端直接复制单词作为译文。此外，对比两个模型在翻译低频词时的正确率。统计结果表明，当扰动概率增加时，BERT-NMT 模型翻译低频词的效果甚至不如普通的 NMT 模

型。可见预训练并不总能为下游任务带来性能上的提升，在使用过程中应该更为小心。

1 研究现状

1.1 NMT 模型中的 PLM

神经机器翻译与预训练的交叉领域主要包含以下两条研究线路：① 将 PLM 融合至神经机器翻译模型；② 预训练跨语言的语言模型。

1.1.1 将 PLM 融合至神经机器翻译模型

在 BERT[1]被提出之后，出现了许多将 BERT 融入机器翻译模型的简单尝试，包括使用 PLM 的输出替代机器翻译模型的嵌入层[21]，以及使用 PLM 的参数对 NMT 模型的编码器进行初始化[22]。Zhu 等[2]设计了 BERT-Enc 和 BERT-解码器（BERT-Dec，BERT-decoder）注意力模块，并且混合了不同模块输出的表示。类似地，APT 框架使用基于层的注意力机制对 BERT 的各层输出进行动态混合[13]。Guo 等[23]使用两个 BERT 分别作为编码器和解码器，在 BERT 不同层之间添加适配器，以同时利用编码器端和解码器端 BERT 包含的信息。Guo 等[24]进一步提出冻结 BERT 参数并添加适配器的方法来加速机器翻译模型的训练过程。Shavarani 等[15]提出更加充分地利用 BERT 包含的语言学信息，而非简单地将其作为词嵌入的替代品。Xu 等[25]使用 145 GB 德语文本训练了一个特制的双语语言模型 BiBERT，来提高模型的翻译效果。

1.1.2 跨语言的语言模型预训练

由于 BERT 在训练过程中仅使用单语语料，并不适合执行机器翻译任务，Conneau 等[26]设计了一个新的翻译语言模型（TLM，translation language model）任务，并使用双语平行语料训练得到 XLM 模型。此外，以 BERT 为代表的单语语言模型往往仅包含编码器，与翻译模型常采用的编码器–解码器架构并不匹配，为此，Song 等[27]提出预训练一个序列到序列架构的模型 MASS。此后，Liu 等[28]设计了更为复杂的预训练任务，并使用 25 种语言训练得到 mBART，使得在翻译任务上进行微调成为可能。Lin 等[29]使用随机对齐替换任务训练了一个普适性更强的多语言翻译模型 mRASP。Pan 等[30]在 mRASP 的基础上引入对比学习和基于对齐的数据增强以提高模型效果。Li 等[31]的研究表明，对序列到序列模型进行预训练的过程中使用双向解码器能显著提高模型的翻译效果。

1.2 NMT 模型的鲁棒性

尽管相比先前的统计机器翻译模型，神经机器翻译模型生成的译文质量有了显著提高，但其仍存在一些不足之处。Belinkov 等[32]发现基于字符的神经机器翻译模型的翻译效果很容易受到文本中噪声的影响。Cheng 等[33]指出同义词替换同样能够降低神经机器翻译模型的翻译效果。为了检验神经机器翻译模型的鲁棒性，之前的研究还尝试在输入中加入笔误，对输入中的字符或单词执行交换、重复以及删除操作，并对模型的相应输出进行评价[34~35]。

为了提高 NMT 模型的鲁棒性，主流的方法为基于对抗学习[32]，即使用人工添加噪声的数据训练 NMT 模型。Sato 等[36]提出向输入添加使得损失增大最多的噪声以增强模型的鲁棒性。Cheng 等[37]设计了一个基于梯度的方法来生成对抗样本。此外，Sennrich 等[38]以及 Michel 和 Neubig[39]说明使用子词表示句子能够帮助机器翻译模型更好地处理扰动。Cheng 等[33]提出了一个新的训练目标，最大化原始输入和扰动过的样本之间的相似性。UniDrop[40]在特征、结构和数据 3 个维度使用丢弃法来提高机器翻译模型的效果。此外，Cheng 等[41]将有监督训练和自监督训练相结合来提高神经机器翻译模型应对代码转换类型干扰的鲁棒性。

本文使用融合 BERT 的神经机器翻译模型来探究 BERT 为神经机器翻译模型带来了哪些影响。之前的研究表明，BERT 能够提高文本分类模型的鲁棒性[16]。但 BERT 是否会影响神经机器翻译模型的鲁棒性还不得而知，这是本文要探究的问题。

2 BERT 改善句法能力

本节通过实验来分析 BERT 如何影响神经机器翻译模型的建模能力，具体来说，使用 5 种不同的探针任务来检测 NMT 模型的编码器和 BERT-NMT 模型的编码器中分别编码了哪些句

法相关的信息。

2.1　实验设置

2.1.1　探针任务

本文实验共使用以下 5 种探针任务。Distance 任务中探针需要依据任意两个单词的向量表示预测它们在依存句法树中之间的距离。Depth 任务则需要探针根据每个单词的向量表示预测其在依存句法树中的深度，即单词和根节点之间的距离。BShift 任务则需要探针根据句子的向量表示判断句中是否存在两个相邻单词的位置被调换了。TreeDepth 任务要求探针预测句子的短语结构树的最大深度。TopConst 任务要求探针对句子的短语结构树最顶层类型进行分类。表 1 提供了每种句法探针任务的示例。

表 1　句法探针任务的示例

Table 1　Examples for each syntactic probing task

任务	示例	标签
Distance	why not just bring up the idea of staying in your current lab with your advisor ? word: idea, current	3
Depth	Budget negotiations between the mayor and the city council are entering high gear . word: mayo	3
BShift	She wondered how time much had passed .	Inverted
TreeDepth	Who knew who would be there ?	10
TopConst	I wanted to start asking questions now , but forced myself to wait .	NP_VP_

单词 w_i 和 w_j 之间的距离 $d_{\mathrm{T}}(w_i,w_j)$ 定义为两者在依存句法树之间边的数目。对于句子 $s=[w_1,w_2,\cdots,w_n]$，使用探针 B 根据每对单词 (w_i,w_j) 对应的表示 (h_i,h_j) 来预测它们之间的距离。

$$d_{\mathrm{B}}(h_i,h_j)=\left(\mathrm{B}(h_i-h_j)\right)^{\mathrm{T}}\left(\mathrm{B}(h_i-h_j)\right) \tag{1}$$

探针 B 的训练目标为

$$\min_{\mathrm{B}}\sum_{s}\frac{1}{n^2}\sum_{i,j=1}^{n}\left|d_T(w_i,w_j)-d_{\mathrm{B}}(h_i,h_j)\right| \tag{2}$$

单词 w_i 的深度 $d_{\mathrm{T}}(w_i)$ 定义为其在依存句法树中与根节点之间的边数。对于句子 $s=[w_1,w_2,\cdots,w_n]$，根据每个单词 w_i 对应的表示 h_i，使用探针 B 来预测它的深度。

$$d_{\mathrm{B}}(h_i)=(\mathrm{B}h_i)^{\mathrm{T}}(\mathrm{B}h_i) \tag{3}$$

探针 B 的训练目标为

$$\min_{\mathrm{B}}\sum_{s}\frac{1}{n}\sum_{i=1}^{n}\left|d_{\mathrm{T}}(w_i)-d_{\mathrm{B}}(h_i)\right| \tag{4}$$

至于 BShift、TreeDepth 和 TopConst 这 3 个任务，使用包含一层隐藏层的多层感知机（MLP，multi-layer perceptron）分类器作为探针，并使用 Sigmoid 作为激活函数。给定句子 s_i 对应的向量表示 $\bar{\boldsymbol{h}}_i$，预测方法为

$$\hat{y}_i=\mathrm{softmax}\left(\mathrm{MLP}(\bar{\boldsymbol{h}}_i)\right) \tag{5}$$

训练目标为

$$\min_{\theta_{\mathrm{MLP}}}-\sum_{i}y_i^{\mathrm{T}}\log(\hat{y}_i) \tag{6}$$

其中，θ_{MLP} 表示 MLP 的全部参数，y_i 为独热码。

对于 Distance 和 Depth 任务，仿照 Hewitt 和 Manning[10]的研究，使用探针的预测值和真实值之间的斯皮尔曼相关系数 ρ 进行评价，对于其他任务则使用准确率作为评价指标。

2.1.2　数据

对于 Distance 和 Depth 任务，使用 STS 12-16[42-46]的数据，并且使用开源工具 Stanza[47] 生成每个句子对应的依存句法树以计算单词之间的距离和每个单词的深度。对全部数据按照 7:2:1 划分为训练集、验证集和测试集，分别包含 36 000、10 000、5 000 条句子。

对于其他探针任务，使用 SentEval[48]提供的数据集。每个任务的训练集包含 100 000 条句子，验证集和测试集的大小均为 10 000。

使用 WMT14 英德数据集训练神经机器翻译模型。在训练之前，首先使用 40 000 次合并操作的字节对编码[38]（BPE，byte-pair encoding）算法对句子进行编码。验证集为 newstest2013，测试集为 newstest2014。

2.1.3　模型

本节实验中使用到的模型包括两类：基础模型和探针模型。基础模型用于生成单词及句子的向量表示，而探针模型则根据这些向量表示执行相应的任务。对于 Distance 和 Depth 任务，使用矩阵 $\boldsymbol{B}\in\mathbb{R}^{d_{\mathrm{model}}\times r}$ 作为探针，其中 d_{model} 为基础模型

生成的向量表示对应的维度，r 为探针的秩。对于其他任务，使用包含一层隐藏层的 MLP 分类器作为探针。

用于生成单词表示的基础模型包括 BERT、NMT 编码器以及 BERT-NMT 编码器。

NMT 模型的整体架构如图 1 所示，包含编码器和解码器。其中，编码器主要包含自注意力和前馈网络模块，解码器则包括掩码自注意力、交叉注意力以及前馈网络 3 个模块。每个模块之后都会进行残差连接和层标准化操作。

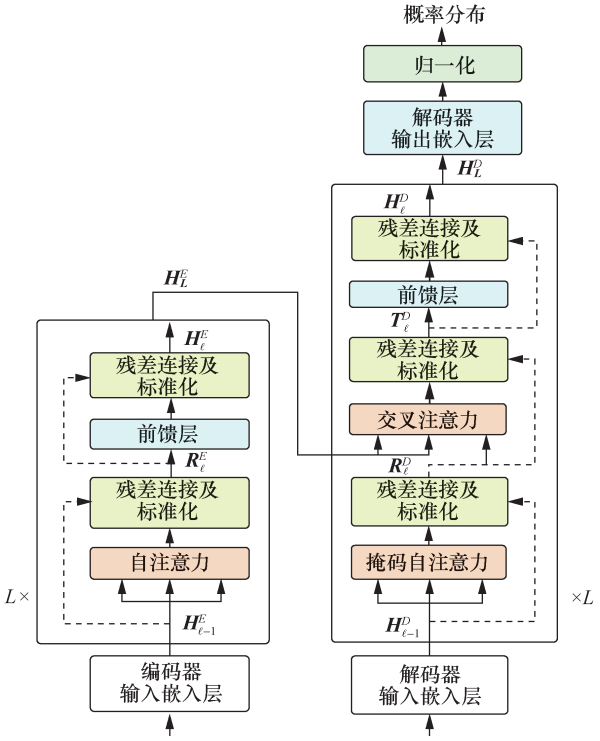


图 1　NMT 模型的整体结构

Figure 1　The whole structure of the NMT model

NMT 模型第 ℓ 层编码器的输出为

$$\boldsymbol{R}_\ell^E = \mathrm{LN}\left(\boldsymbol{H}_{\ell-1}^E + \mathrm{Attn}(\boldsymbol{H}_{\ell-1}^E, \boldsymbol{H}_{\ell-1}^E, \boldsymbol{H}_{\ell-1}^E)\right) \tag{7}$$

$$\boldsymbol{H}_\ell^E = \mathrm{LN}\left(\boldsymbol{R}_\ell^E + \mathrm{FFN}(\boldsymbol{R}_\ell^E)\right) \tag{8}$$

其中，$\mathrm{LN}(\cdot)$ 表示层标准化操作。

注意力模块 $\mathrm{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 的计算方法为

$$\mathrm{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}^Q)\cdot(\boldsymbol{K}\boldsymbol{W}^K)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\boldsymbol{W}^V \tag{9}$$

其中，$\boldsymbol{W}^Q,\boldsymbol{W}^K \in \mathbb{R}^{d_{\mathrm{model}}\times d_k}$，$\boldsymbol{W}^V \in \mathbb{R}^{d_{\mathrm{model}}\times d_v}$ 为可训练参数。

前馈网络 $\mathrm{FFN}(h)$ 的计算方法为

$$\mathrm{FFN}(h) = \mathrm{ReLU}(h\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2 \tag{10}$$

其中，$\boldsymbol{W}_1,\boldsymbol{W}_2 \in \mathbb{R}^{d_{\mathrm{model}}\times d_{\mathrm{model}}}$，$b_1,b_2 \in \mathbb{R}^{d_{\mathrm{model}}}$ 为可训练参数。

类似地，NMT 模型第 ℓ 层解码器的输出可以表示为

$$\boldsymbol{R}_\ell^D = \mathrm{LN}\left(\boldsymbol{H}_{\ell-1}^D + \mathrm{Attn}(\boldsymbol{H}_{\ell-1}^D, \boldsymbol{H}_{\ell-1}^D, \boldsymbol{H}_{\ell-1}^D)\right) \tag{11}$$

$$\boldsymbol{T}_\ell^D = \mathrm{LN}\left(\boldsymbol{R}_\ell^D + \mathrm{Attn}(\boldsymbol{R}_\ell^D, \boldsymbol{H}_L^E, \boldsymbol{H}_L^E)\right) \tag{12}$$

$$\boldsymbol{H}_\ell^D = \mathrm{LN}\left(\boldsymbol{T}_\ell^D + \mathrm{FFN}(\boldsymbol{T}_\ell^D)\right) \tag{13}$$

BERT-NMT 模型由 BERT-fused 的编码器和 Transformer 模型的解码器构成。BERT-fused 的编码器结构如图 2(a)所示，通过 BERT-Enc 注意力模块将 BERT 最后一层的输出 $\boldsymbol{H}_{L_{\mathrm{BERT}}}^B$ 引入 NMT 模型，其第 ℓ 层的输出可以写为

$$\boldsymbol{R}_\ell^E = \mathrm{LN}\left(\begin{array}{l}\boldsymbol{H}_{\ell-1}^E + \gamma_\ell\,\mathrm{Attn}(\boldsymbol{H}_{\ell-1}^E, \boldsymbol{H}_{\ell-1}^E, \boldsymbol{H}_{\ell-1}^E) + \\ (1-\gamma_\ell)\,\mathrm{Attn}(\boldsymbol{H}_{\ell-1}^E, \boldsymbol{H}_{L_{\mathrm{BERT}}}^B, \boldsymbol{H}_{L_{\mathrm{BERT}}}^B)\end{array}\right) \tag{14}$$

$$\boldsymbol{H}_\ell^E = \mathrm{LN}\left(\boldsymbol{R}_\ell^E + \mathrm{FFN}(\boldsymbol{R}_\ell^E)\right) \tag{15}$$

其中，$\gamma_\ell = 0.5, \forall \ell \in [1,6]$。

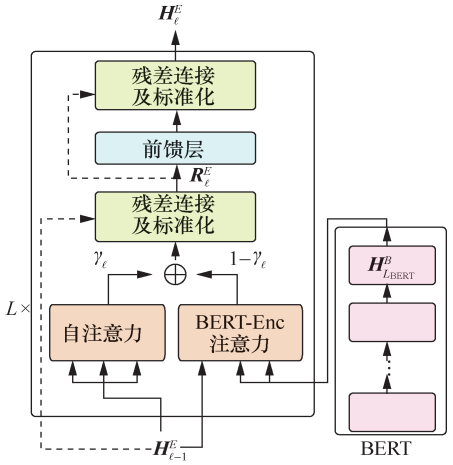
对于 BERT，使用 HuggingFace[49]提供的 bert-base-uncased 模型作为基准。NMT 模型和 BERT-NMT 模型的实现与训练均基于 Fairseq[50]，其中 BERT-NMT 模型在训练时同样采用 bert-base-uncased 作为 BERT 模块。

对于每个基础模型，使用其最后一层的输出作为单词的表示，并且使用单词表示的均值作为句子的嵌入。

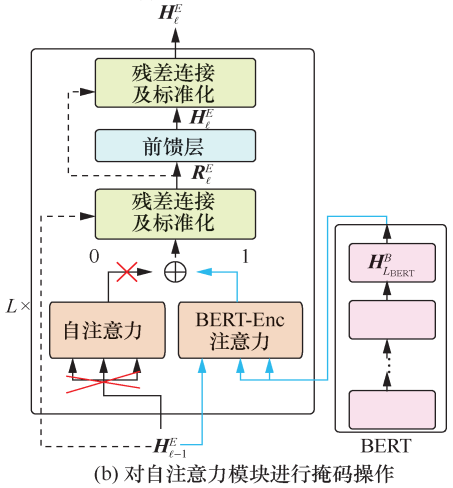
此外，为了能够了解自注意力模块和 BERT-Enc 模块的作用，使用控制变量法，在训练结束后生成单词表示时分别对两者进行掩码操作，这样最终得到的向量表示仅由一部分模块计算得到。对自注意力模块进行掩码操作如图 2(b)所示，对 BERT-Enc 模块进行掩码操作图 2(c)所示。

2.1.4　训练设置

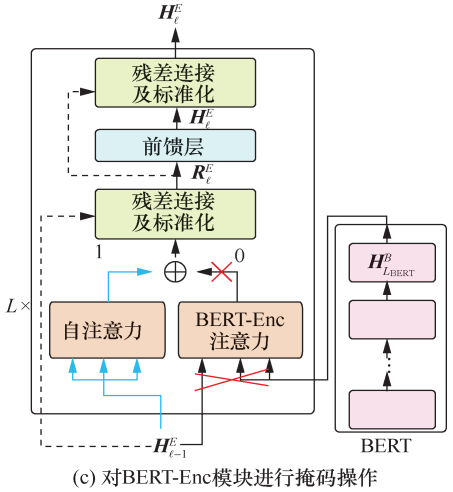
对于 Distance 和 Depth 探针，秩 r 设置为 64，当基础模型为翻译模型时，设置 $d_{\mathrm{model}} = 512$；当基础模型为 BERT 时，则设为 768。探针的训练目标为最小化预测值和实际值的 L1 损失。优化器使用



(a) BERT-fused的编码器结构



(b) 对自注意力模块进行掩码操作



(c) 对BERT-Enc模块进行掩码操作

图 2　BERT-fused 编码器结构和掩码操作示意
Figure 2　Overview of the structure of BERT-fused encoder and masking methods

Adam[51]，其中 $\beta_1=0.9,\beta_2=0.999,\varepsilon=1\times10^{-8}$。学习率设置为 0.001。探针被训练直至收敛，最多训练 40 轮，批大小为 20。至于其他探针，其训练目标为最小化预测值和实际值之间的交叉熵损失。学习率同样设置为 0.001。优化器依然采用 Adam，但是参数设置略有不同：$\beta_1=0.9,\beta_2=0.98$。类似地，探针最多训练 200 轮直至收敛，批大小为 64。

NMT 和 BERT-NMT 模型的隐藏层维度设置为 512，注意力头个数为 8，前馈层维度为 2 048。学习率为 0.000 7，训练时先预热 4 000 步。将丢弃率设置为 0.1，标签平滑使用参数 $\varepsilon=0.1$。优化器 Adam 的参数为 $\beta_1=0.9,\beta_2=0.98$。全部实验在两张 NVIDIA GTX 1080Ti GPU 上完成，分配给每张显卡的训练批次中包含 4 096 个子词。使用梯度累积来模拟更大的批大小。翻译模型训练到在验证集上收敛为止。

2.1.5　翻译模型结果

本文使用 SacreBLEU[52]计算模型生成译文的 BLEU 分数，NMT 模型分数为 27.1，BERT-NMT 模型为 29.0，这与之前的文献结果一致[2,17]，也保证了之后对比分析的有效性。

2.2　BERT 的影响

句法相关的探针任务实验结果如表 2 所示，在 BERT 的帮助下，BERT-NMT 编码器在 Distance、Depth、BShift、TreeDepth、TopConst 任务上都获得了比 NMT 编码器更高的准确率。

此外，对自注意力模块和 BERT-Enc 注意力模块分别进行掩码操作会导致准确率发生不同变化。虽然对 BERT-NMT 编码器中的自注意力模块进行掩码操作会导致模型在 Distance 和 Depth 任务上的性能出现略微下降，掩盖掉 BERT-Enc 注意力模块则会导致模型在 BShift 任务上的准确率大幅下跌 18.51%。作为对比，对自注意力模块进行掩盖甚至会导致探针的准确率上涨 0.83%。考虑到 BERT 模型本身在 BShift 任务上能够取得较高的准确率，BERT 提高了神经机器翻译模型根据语义识别有问题的词序的能力。这说明 BERT 能够从混乱的语序中提取语义信息，因此下文讨论 BERT 能否提高机器翻译模型的鲁棒性。

3　使用语序扰动攻击 NMT 模型

即使相邻的两个单词位置调换，BERT 也能够识别出句子的语义信息，这种特性可以帮助神经机器翻译模型更好地应对源语言句子中的扰

表 2　句法相关的探针任务实验结果
Table 2　Results of syntactic probing tasks

模型	准确率				
	Distance	Depth	BShift	TreeDepth	TopConst
BERT	74.16%	78.79%	88.77%	36.21%	72.62%
NMT 编码器	80.86%	83.53%	70.66%	39.94%	75.26%
BERT-NMT 编码器	81.09%	83.97%	85.14%	40.66%	70.27%
掩码自注意力模块	77.91%	79.96%	85.97%	38.75%	71.74%
掩码 BERT-Enc 注意力模块	80.22%	82.07%	66.63%	39.80%	74.04%

动。本节探究 BERT 是否能够提升机器翻译模型的鲁棒性。

3.1　实验设置

3.1.1　攻击方法

为了探究 BERT 能否提高机器翻译模型的鲁棒性，本节仿照 BShift 任务设计了一种扰乱词序的攻击方法。简单来说，打乱翻译任务测试集中句子的原有词序，并以一定概率 p 交换两个相邻单词，之后以扰动后的句子作为输入，评估模型翻译出的译文效果。表 3 展示了使用攻击方法按照不同概率生成的样例。值得注意的是，句首第一个单词和结尾标点的位置在攻击时不会被改变。

表 3　使用攻击方法按照不同概率生成的样例
Table 3　Examples generated by attacking method according to different probabilities

p	句子
0.0	Two sets of lights so close to one another: intentional or just a silly error?
0.1	Two sets of lights so close to one another: intentional or just silly a error?
0.2	Two sets of so lights close to one another: or intentional a just silly error?
0.3	Two of sets lights so to close another one intentional or: just a silly error?
0.4	Two sets of so lights to one close: another or intentional just a silly error?
0.5	Two sets lights of close to so one another intentional: or just silly a error?

该攻击方法的形式化定义如下：假设 s 是一个由 n 个单词组成的句子，即 $s=[w_1,w_2,\cdots,w_n]$。对于每个位置 $i,i\in[2,n-2]$，w_i 以概率 p 与自己右边相邻的单词 w_{i+1} 交换位置。这样，s 将会变成 $[w_1,\cdots,w_{i+1},w_i,\cdots,w_n]$。使用 s^p 表示依照概率 p 对 s 进行扰动得到的结果，其中 $p\in(0,0.5]$。

攻击之后，分别观察 NMT 和 BERT-NMT 模型输出的翻译结果。如果 BERT 能够提高机器翻译模型的鲁棒性，那么 BERT-NMT 的译文分数（BLEUScore 或 BERTScore）将会更高。反之，如果 BERT-NMT 模型的分数更低，则说明 BERT 可能对机器翻译模型的鲁棒性有负面影响。

3.1.2　数据

为了囊括尽可能多的语言，本文选择了 5 种来自不同语系的语言对，包括 WMT14 德语−英语数据集（De-En，包括 4 500 000 条平行语料）、WMT14 英语−德语数据集（En-De）、WMT17 芬兰语−英语数据集（Fi-En，包括 2 600 000 条平行语料）、WMT17 土耳其语−英语数据集（Tr-En，包括 207 000 条平行语料）以及 WMT17 中文−英语数据集（Zh-En，包括 20 800 000 条平行语料）。对于 WMT14 英德数据集，借助 Moses 提供的脚本进行分词，之后使用 40 000 次 BPE 合并操作进行编码，并生成一个共享字典。对于 WMT17 Fi-En 和 Tr-En，使用 WMT17 官方提供的预处理后的版本，同样使用源语言和目标语言的共享字典。至于 WMT17 Zh-En 数据集，首先使用 jieba 对中文文本进行分词处理，之后对中文和英文数据集分别使用 32 000 次 BPE 合并操作构建字典。

对于 WMT14 数据集，使用 newstest2013 作为验证集，newstest2014 作为测试集。WMT17 的翻译任务在 newstest2016 上进行验证，在 newstest2017 上进行评测。

3.1.3　模型

本节使用的 NMT 模型和 BERT-NMT 模型的架构与超参数与 2.1 节中介绍的相同。

表 4　NMT 和 BERT-NMT 在相应测试集上的 BLEUScore 和 BERTScore
Table 4　BLEUScore and BERTScore for NMT and BERT-NMT on corresponding test sets

模型	De-En BLEUScore/ BERTScore	En-De BLEUScore/ BERTScore	Fi-En BLEUScore/ BERTScore	Tr-En BLEUScore/ BERTScore	Zh-En BLEUScore/ BERTScore
NMT	31.1 / 85.8	27.1 / 85.5	25.9 / 83.8	16.0 / 75.5	22.8 / 82.4
BERT-NMT	32.4 / 86.8	29.0 / 86.1	26.9 / 84.9	18.8 / 78.6	23.2 / 82.7

3.1.4　评价指标

使用 BLEUScore 和 BERTScore 来评价翻译结果。根据 Zhang 等[20]的研究，相比基于 n 元语法的评价方法，BERTScore 能够惩罚与语义有关的词序变化，比较适合用于评价语义变化。本文采用 BERTScore 的 $F1$ 值作为评价指标。表 4 展示了 NMT 和 BERT-NMT 在相应测试集上的 BLEUScore 和 BERTScore。尽管语言对不同，BERT-NMT 模型在 BLEUScore 和 BERTScore 上都明显高于 NMT 模型，说明将 BERT 融合到机器翻译模型中确实能够提高模型生成的译文质量。

3.2　攻击结果

3.2.1　BLEUScore 和 BERTScore 的变化

图 3 分别展示了 BERT-NMT 和 NMT 在 BLEUScore 和 BERTScore 上的差值。尽管语言对之间的亲疏关系不同，但除了 En-De 任务之外，大多具有相近的趋势。对于其他语言来说，不论攻击概率 p 如何变化，BERT-NMT 模型都能将译文的 BLEUScore 和 BERTScore 提高1.0 左右。这归功于 BERT 模型从看似无序的单词中提取语义信息的能力。值得注意的是，BLEUScore 和 BERTScore 的变化都证实英文 BERT 会给 NMT 模型的翻译结果带来负面影响：当攻击概率 p 大于30% 时，BERT-NMT 模型产生的译文比 NMT 模型生成的译文更差。

3.2.2　源语言端和目标语言端语义的变化

为了找出融合英文 BERT 后，BERT-NMT 模型翻译质量下降这一现象背后的原因，比较源语言端句子和目标语言端译文的语义变化。

$$\Delta^{p}_{\text{src}} = \frac{1}{n}\sum_{i}\left|\text{BS}(s_i, s_i^{p}) - \text{BS}(s_i, s_i)\right| \tag{16}$$

$$\Delta^{p}_{\text{tgt}} = \frac{1}{n}\sum_{i}\left|\text{BS}(r_i, t_i^{p}) - \text{BS}(r_i, t_i)\right| \tag{17}$$

其中，$\text{tgt} \in \{\text{NMT}, \text{BERT-NMT}\}$。$s_i^{p}$ 表示按照概率 p 对句子 s_i 进行扰动的结果。t_i 表示神经机器翻译模型根据输入 s_i 生成的译文，而 r_i 则表示 s_i 对应的标准翻译。$\text{BS}(s_i, s_j)$ 为 s_j 相对于 s_i 的 BERTScore。

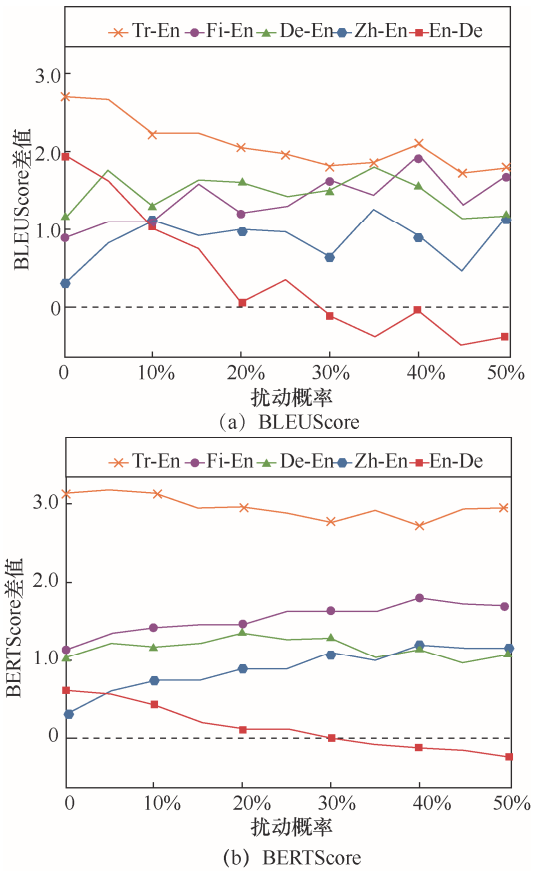


图 3　BERT-NMT 和 NMT 在 BLEUScore 和 BERTScore 上的差值
Figure 3　Plots of differences between BERT-NMT and NMT in terms of BLEUScore and BERTScore

由于对不同语言计算 BERTScore 时使用了不同的语言模型，直接对 BERTScore 进行比较并不合理，因此使用累积分布函数（CDF，cumulative distribution function）将得到的分数转化为[0,1]上的分布，实现对 BERTScore 的归一化。之后若没有特别指出，将使用 $\text{BS}(s_i, s_j)$ 来代表经 CDF 变换后的 BERTScore。

源语言和目标语言端语义变化如图 4 所示，

En-De 在源语言端的语义变化明显大于 De-En 中源语言端的语义变化，说明不同语言对于词序扰动具有不同的反应。对于英语文本来说，受到干扰之后句子的语义发生了一定改变，而德语的句子仍保留了原始的意义。一种可能的解释是德语的语法允许较为灵活的词序。

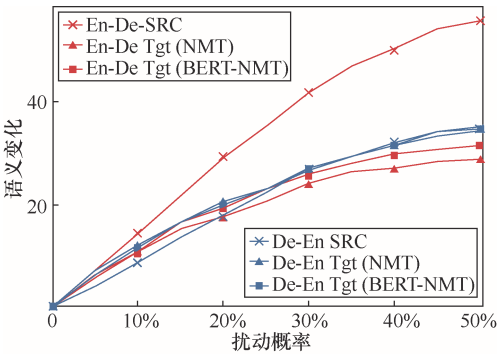


图 4　源语言和目标语言端语义变化
Figure 4　Changes of semantics in the source side and target side

但是在 En-De 任务中，NMT 和 BERT-NMT 模型在目标语言端上的变化趋势和源语言端并不相同。而它们在 De-En 任务中的变化趋势和源语言端的变化更为类似。这说明在 En-De 任务中，不论 NMT 还是 BERT-NMT 模型，都没能很好地应对攻击带来的变化。

3.2.3　源语言端和目标语言端语义的差值

为了更好地描绘源语言端和目标语言端的变化，定义两端之间语义的差值 Diff 。

$$\mathrm{Diff}_{\mathrm{tgt}}^{p}=\frac{1}{n}\sum_{i}\left|\left|\mathrm{BS}(s_i,s_i^p)-\mathrm{BS}(s_i,s_i)\right|-\left|\mathrm{BS}(r_i,t_i^p)-\mathrm{BS}(r_i,t_i)\right|\right| \tag{18}$$

其中，$\mathrm{tgt}\in\{\mathrm{NMT},\mathrm{BERT\text{-}NMT}\}$，$p$ 为扰动概率。

在 De-En 和 En-De 任务上一同比较 Diff 是合理的，因为它们涉及的语言是相同的，而且在计算过程中使用了相同的模型进行打分。

图 5 展示了源语言和目标语言端间的语义差值。当攻击的扰动概率 p 增加时，BERT-NMT 模型在 En-De 上的 Diff 明显高于 De-En 上的 Diff 值，说明将英语 BERT 融合至神经机器翻译模型中难以缩小语义差别。因此，英文 BERT 并没有为机器翻译模型抵御源语言句子中扰动的鲁棒性带来太多提升。

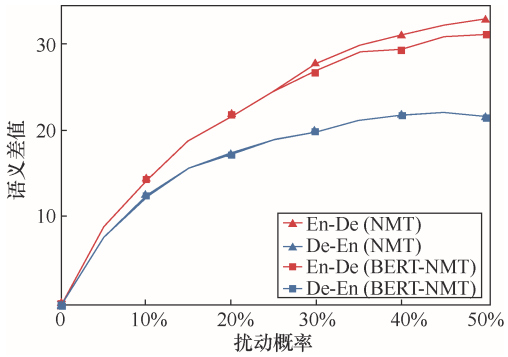


图 5　源语言和目标语言端间的语义差值
Figure 5　Semantic difference between the source side and target side

3.3　分析

3.3.1　复制行为

为了更加细致地了解翻译质量下降的表现，本文比较了 NMT 模型和 BERT-NMT 模型在输入未受扰动和受扰动情况下生成的译文。当输入分别为未受攻击的句子和扰动后的句子时 NMT 和 BERT-NMT 生成的翻译样例如表 5 所示，当源语言句子未被攻击时，BERT-NMT 模型能够产生高质量的翻译。但当输入受到扰动后，BERT-NMT 模型倾向于简单地直接从源语言句子中复制单词。而 NMT 模型仍能生成较为合理的译文。统计结果显示，BERT-NMT 模型从源语言端错误地复制了 48 个句子，而 NMT 模型仅复制了 33 个句子。这一对比结果说明，随着攻击概率的增加，融合 BERT 将会放大神经机器翻译模型由于数据不确定性引发的复制错误[5,53]。

3.3.2　低频词翻译不足

考虑到翻译低频词时往往需要依赖上下文信息，而前文提出的攻击方式会破坏上下文信息，因此 En-De 任务中 BERT-NMT 模型翻译效果下降的一个可能解释是模型难以很好地处理低频词的翻译。compare-mt[54]能够用于评价单词级别的翻译表现。根据单词在训练集中出现的频率将其划分为 3 类：低频词（小于 10）、中频词（10 到 100）和高频词（大于 100）。对于每一类，compare-mt 报告预测结果相对真实结果的 $F1$ 值。Tr-En 和 En-De 中不同频率单词 NMT 和 BERT-NMT 的翻译 $F1$ 值如表 6 和表 7 所示。不论词频和攻击概率如何变化，使用 Tr-En 数据集训练得到的 BERT-NMT 模型的 $F1$ 值始终高于 NMT 模型。

表 5　当输入分别为未受攻击的句子和扰动后的句子时 NMT 和 BERT-NMT 生成的翻译样例

Table 5　Translation examples generated by NMT and BERT-NMT when presented with undamaged and perturbed source sentence

条件	句子
源语言句子	In Cameroon, there is only one doctor for every 5 000 people, according to the World Health Organization.
参考翻译	In Kamerun gibt es nur einen Arzt für je 5 000 Menschen, so die Weltgesundheitsorganisation.
NMT	Nach Angaben der Weltgesundheitsorganisation gibt es in Kamerun nur einen Arzt für jeden 5 000 Menschen.
BERT-NMT	In Kamerun gibt es laut Weltgesundheitsorganisation nur einen Arzt pro 5 000 Menschen.
源语言句子（p=0.5）	In, there is only one Cameroon for every 5 000 people, according to doctor the Health Organization World.
NMT（p=0.5）	In, gibt es nur ein Kamerun für jeden 5 000 Menschen, so der Arzt der Weltgesundheitsorganisation World.
BERT-NMT（p=0.5）	In, there is only one Cameroon for each 5 000 people, according to doctor the Health Organization World.

表 6　Tr-En 中不同频率单词 NMT 和 BERT-NMT 的翻译 $F1$ 值

Table 6　Word prediction $F1$ score of NMT and BERT-NMT on words with different frequencies in Tr-En

模型（攻击概率）	$F1$ 值		
	低频词	中频词	高频词
NMT（p=0）	24.1	28.4	46.8
BERT-NMT（p=0）	27.1	32.3	50.7
NMT（p=0.5）	19.0	22.9	41.7
BERT-NMT（p=0.5）	19.4	24.9	44.5

表 7　En-De 中不同频率单词 NMT 和 BERT-NMT 的翻译 $F1$ 值

Table 7　Word prediction $F1$ score of NMT and BERT-NMT on words with different frequencies in En-De

模型（攻击概率）	$F1$ 值		
	低频词	中频词	高频词
NMT（p=0）	31.6	40.5	53.0
BERT-NMT（p=0）	33.5	43.2	54.9
NMT（p=0.5）	20.1	26.1	40.7
BERT-NMT（p=0.5）	19.8	26.4	41.1

当攻击概率为0.5时，使用 En-De 数据集训练得到的 BERT-NMT 模型在翻译低频词时比 NMT 模型差。当文本未受到攻击时，BERT-NMT 模型在翻译低频词时的 $F1$ 值比 NMT 模型的高1.9。而当扰动概率为0.5时，BERT-NMT 模型在低频词上的 $F1$ 值仅为19.8，甚至低于 NMT 模型的20.1。这说明加入英文 BERT 后，机器翻译模型鲁棒性受损的另一个表现：当词序扰动概率增加时，BERT-NMT 模型不能很好地处理低频词。

4　结束语

本文说明了 BERT 能够提高神经机器翻译模型句法相关的能力，特别是对词序进行建模的能力；设计了一种基于词序扰动的攻击方法，来检测翻译模型的鲁棒性。在多种不同语言对上的实验结果表明，融合英语 BERT 会给神经机器翻译模型的翻译质量带来消极影响。英语 BERT 无法缩小原始句子和受扰动句子之间的语义差别。进一步分析显示，BERT 导致了更多翻译时的复制行为，以及低频词翻译不足的后果。因此，研究者应该根据下游任务和具体情境决定是否使用预训练语言模型。未来，将继续探究跨语言预训练方法对神经机器翻译模型语义、句法能力以及鲁棒性的影响。

参考文献：

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

[2] ZHU J, XIA Y, WU L, et al. Incorporating BERT into neural machine translation[C]// 8th International Conference on Learning Representations. 2020: 1-16.

[3] BAZIOTIS C, HADDOW B, BIRCH A. Language model prior for low-resource neural machine translation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7622-7634.

[4] AN T, SONG J, LIU W. Incorporating pre-trained model into neural machine translation[C]//2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD). 2021: 212-216.

[5] LIU X, WANG L, WONG D F, et al. On the complementarity between pre-training and back-translation for neural machine translation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 2900-2907.

[6] PETERS M E, NEUMANN M, ZETTLEMOYER L, et al. Dissecting contextual word embeddings: architecture and representation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1499-1509.

[7] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn

about the structure of language[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3651-3657.

[8] TENNEY I, DAS D, PAVLICK E. BERT rediscovers the classical NLP pipeline[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4593-4601.

[9] TENNEY I, XIA P, CHEN B, et al. What do you learn from context? Probing for sentence structure in contextualized word representations[C]//7th International Conference on Learning Representations. 2019: 1-17.

[10] HEWITT J, MANNING C D. A structural probe for finding syntax in word representations[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4129-4138.

[11] GOLDBERG Y. Assessing BERT's syntactic abilities[J]. arXiv preprint arXiv:1901.05287, 2019.

[12] SUNDARARAMAN D, SUBRAMANIAN V, WANG G, et al. Syntax-infused transformer and bert models for machine translation and natural language understanding[J]. arXiv preprint arXiv:1911.06156, 2019.

[13] WENG R, YU H, HUANG S, et al. Acquiring knowledge from pre-trained model to neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9266-9273.

[14] YANG J, WANG M, ZHOU H, et al. Towards making the most of bert in neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9378-9385.

[15] SHAVARANI H S, SARKAR A. Better neural machine translation by extracting linguistic information from BERT[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021: 2772-2783.

[16] HAUSER J, MENG Z, PASCUAL D, et al. BERT is robust! a case against synonym-based adversarial examples in text classification[J]. arXiv preprint arXiv: 2109.07403, 2021.

[17] 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.

[18] CONNEAU A, KRUSZEWSKI G, LAMPLE G, et al. What you can cram into a single $ &!#* vector: probing sentence embeddings for linguistic properties[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 2126-2136.

[19] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[20] ZHANG T, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT[C]//8th International Conference on Learning Representations. 2020: 1-43.

[21] CLINCHANT S, JUNG K W, NIKOULINA V. On the use of BERT for neural machine translation[C]//Proceedings of the 3rd Workshop on Neural Generation and Translation. 2019: 108-117.

[22] ROTHE S, NARAYAN S, SEVERYN A. Leveraging pre-trained checkpoints for sequence generation tasks[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 264-280.

[23] GUO J, ZHANG Z, XU L, et al. Incorporating bert into parallel sequence decoding with adapters[J]. Advances in Neural Information Processing Systems. 2020, 33: 10843-10854.

[24] GUO J, ZHANG Z, XU L, et al. Adaptive adapters: an efficient way to incorporate BERT into neural machine translation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1740-1751.

[25] XU H, VAN DURME B, MURRAY K. BERT, mBERT, or BiBERT? a study on Contextualized Embeddings for Neural Machine Translation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 6663-6675.

[26] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 7059-7069.

[27] SONG K, TAN X, QIN T, et al. MASS: masked sequence to sequence pre-training for language generation[C]//International Conference on Machine Learning. 2019: 5926-5936.

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

[29] LIN Z, PAN X, WANG M, et al. Pre-training multilingual neural machine translation by leveraging alignment information[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 2649-2663.

[30] PAN X, WANG M, WU L, et al. Contrastive learning for many-to-many multilingual 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. 2021: 244-258.

[31] LI P, LI L, ZHANG M, et al. Universal conditional masked language pre-training for neural machine translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 6379-6391.

[32] BELINKOV Y, BISK Y. Synthetic and natural noise both break neural machine translation[C]//6th International Conference on Learning Representations. 2018: 1-13.

[33] CHENG Y, TU Z, MENG F, et al. Towards robust neural machine translation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 1756-1766.

[34] VAIBHAV V, SINGH S, STEWART C, et al. Improving robustness of machine translation with synthetic noise[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 1916-1920.

[35] MICHEL P, LI X, NEUBIG G, et al. On evaluation of adversarial perturbations for sequence-to-sequence models[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 3103-3114.

[36] SATO M, SUZUKI J, KIYONO S. Effective adversarial regularization for neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 204-210.

[37] CHENG Y, JIANG L, MACHEREY W, et al. AdvAug: robust ad-

versarial augmentation for neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5961-5970.

[38] 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. 2016: 1715-1725.

[39] MICHEL P, NEUBIG G. MTNT: a testbed for machine translation of noisy text[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 543-553.

[40] WU Z, WU L, MENG Q, et al. UniDrop: a simple yet effective technique to improve transformer without extra cost[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 3865-3878.

[41] CHENG Y, WANG W, JIANG L, et al. Self-supervised and supervised joint training for resource-rich machine translation[C]//International Conference on Machine Learning. 2021: 1825-1835.

[42] AGIRRE E, CER D, DIAB M, et al. SemEval-2012 task 6: A pilot on semantic textual similarity[C]//Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). 2012: 385-393.

[43] AGIRRE E, CER D, DIAB M, et al. * SEM 2013 shared task: semantic textual similarity[C]//Proceedings of the Main Conference and the shared task: Semantic Textual Similarity. 2013: 32-43.

[44] AGIRRE E, BANEA C, CARDIE C, et al. Multilingual semantic Textual Similarity[C]//Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014: 81-91.

[45] AGIRRE E, BANEA C, CARDIE C, et al. Semantic textual similarity, english, spanish and pilot on interpretability[C]//Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015: 252-263.

[46] AGIRRE E, BANEA C, CER D, et al. Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation[C]//Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 2016: 497-511.

[47] QI P, ZHANG Y, ZHANG Y, et al. Stanza: a python natural language processing toolkit for many human languages[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2020: 101-108.

[48] CONNEAU A, KIELA D. SentEval: an evaluation toolkit for universal sentence representations[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018: 1699-1704.

[49] WOLF T, DEBUT L, SANH V, et al. Transformers: state-of-the-art natural language processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020: 38-45.

[50] OTT M, EDUNOV S, BAEVSKI A, et al. Fairseq: a fast, extensible toolkit for sequence modeling[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). 2019: 48-53.

[51] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]// 3rd International Conference on Learning Representations. 2015: 1-15.

[52] POST M. A call for clarity in reporting BLEU scores[C]//Proceedings of the Third Conference on Machine Translation: Research Papers. 2018: 186-191.

[53] OTT M, AULI M, GRANGIER D, et al. Analyzing uncertainty in neural machine translation[C]//International Conference on Machine Learning. 2018: 3956-3965.

[54] NEUBIG G, DOU Z Y, HU J, et al. compare-mt: a tool for holistic comparison of language generation systems[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). 2019: 35-41.

[作者简介]



赵彧然（1998–），男，河南安阳人，上海交通大学硕士生，主要研究方向为自然语言处理。



薛傥（1999–），男，山西运城人，上海交通大学硕士生，主要研究方向为自然语言处理。



刘功申（1974–），男，山东聊城人，上海交通大学教授、博士生导师，主要研究方向为人工智能安全、自然语言处理、信息安全。