李蔓菁, 迟春诚, 李付学, 等. 基于多译文神经机器翻译数据增强方法[J]. 智能计算机与应用,2024,14(6):35-40. DOI:10.20169/j.issn.2095-2163.240605

基于多译文神经机器翻译数据增强方法

李蔓菁[1], 迟春诚[1], 李付学[2], 闫 红[2]
(1 沈阳化工大学 计算机科学与技术学院, 沈阳 110142; 2 营口理工学院 电气工程学院, 辽宁 营口 115014)

摘 要: 神经机器翻译(NMT)是目前机器翻译领域的主流技术之一,然而其翻译性能的优劣很大程度上取决于数据集的规模和质量。为了缓解数据集稀缺的问题,本文提出了一种基于多译文神经机器翻译的数据增强方法。首先,利用已训练的神经机器翻译模型翻译出多译文,接着,利用筛选策略选出多个译文,同时提出生成伪双语数据的 2 种策略:根据筛选出的译文找到对应的目标原文;利用反向翻译模型对多译文翻译。最后,生成的伪数据与原数据混合,训练增强后的翻译模型。实验结果表明,基于多译文数据增强方法可以有效提高 NMT 模型的翻译性能。

关键词: 神经机器翻译; 数据增强; 多译文; 反向翻译
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2024)06-0035-06

A multi-translation-based data augmentation method for low-resource neural machine translation

LI Manjing[1], CHI Chuncheng[1], LI Fuxue[2], YAN Hong[2]
(1 College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China; 2 College of Electrical Engineering, Yingkou Institute of Technology, Yingkou 115014, Liaoning, China)

Abstract: Neural Machine Translation (NMT) is one of the mainstream technologies in the field of machine translation today, but its translation performance depends largely on the size and quality of the dataset. In order to alleviate the problem of dataset scarcity, this paper proposes a data augmentation method based on multi-translation neural machine translation. Firstly, the trained neural machine translation model is used to translate multiple translations; Then, the multi-translation screening strategy is used to select pseudo-monolingual data, and the target translation is generated using two strategies, which are finding the target text based on the filtered translations and translating multiple translations using a reverse translation model. Finally, the generated pseudo data is mixed with the original data to train an enhanced translation model. Experimental results show that the multi-translation data augmentation method can effectively improve the translation performance of NMT model.

Key words: neural machine translation; data augmentation; multi-translation; back translation

0 引 言

自 20 世纪 40 年代以来,机器翻译一直是自然语言处理领域的关键任务之一。其发展经历了基于规则和基于统计的机器翻译,直到近年来的神经机器翻译(Neural Machine Translation,NMT)。神经机器翻译模型是一种端到端的架构,其中编码器负责将源语言序列编码为一个固定长度的向量,解码器负责将向量解码为目标语言序列。2017 年,Vaswani 等学者[1]提出一种基于 Transformer 的 NMT 模型。Transformer 完全摒弃了递归和卷积,而是基于自注意力和位置编码来捕捉序列中的依赖关系。Transformer 以其并行性和高效性在翻译任务中取得了显著的性能提升。

相比较传统的基于规则和基于统计的机器翻译,神经机器翻译具有更强的语言建模能力和上下文理解能力,而这在先前的工作中已得到证明[2-3]。然而,神经机器翻译的性能仍然受限于训练数据的质量和数量。当神经机器翻译面临稀缺资源时,会因数据稀少而导致翻译不佳。为了解决该问题,研

基金项目: 辽宁省自然科学基金(2021-YKLH-12, 2022-YKLH-18)。
作者简介: 李蔓菁(1999-),女,硕士研究生,主要研究方向:神经机器翻译。
通讯作者: 闫 红(1984-),女,副教授,主要研究方向:神经机器翻译。Email: yanhong@ yku.edu.cn
收稿日期: 2023-05-05

究人员提出多种数据增强方法用于扩充训练数据集。总地来说,可分为 2 类:基于词级的数据增强和基于句子级的数据增强。对此可展开分述如下。

(1)基于词级的方法。是从原始的双语平行语料选择句子中的词通过替换、删除等方法生成新的伪数据,以增加训练数据的数量和多样性,从而提高神经机器翻译模型的性能。具体而言,Jason 等学者[4]使用 WordNet 作为近义词表,从句子中随机选择 N 个非停用词,使用词表中近义词随机替换,进而扩充了语料,缓解了一词多义的问题。Wang 等学者[5]提出了 switchout 方法,利用已有双语训练语料形成源端-目标端词表,利用词表对源端和目标端进行同义词的替换。对于稀缺资源中的稀有词翻译不佳问题,Fadaee 等学者[6]提出用高频词替换低频词,提高了稀有词在机器翻译中的正确率。为了解决替换词性不匹配问题, Muhammad 等学者[7]提出对每个词添加词性标签,利用近义词表计算余弦相似度获得替换相同词性的词。此外, Wu 等学者[8]利用掩码语言模型将预测的单词替换被遮掩的单词,以增强数据的多样性。

(2)基于句子级的方法。最常用的是回译的方法,增加数据的丰富性和多样性,提高翻译模型的性能。Sennrich 等学者[9]提出利用反向翻译模型来翻译单语数据,生成源语言端的译文,将生成的伪数据同原始数据混合,提升数据的丰富性。Zhang 等学者[10]利用反向翻译后的数据再次进行正向翻译,生成更加平滑的伪双语数据。Hoang 等学者[11]认为优质的回译系统可以产生更流畅的双语数据,故通过迭代回译的策略训练出更佳的翻译模型。

本文对句子级的数据增强做了进一步研究,提出一种基于多译文的神经机器翻译数据增强方法。具体而言,本方法利用训练好的翻译模型进行原始数据翻译,翻译出的多译文与参考译文比较,筛选出翻译不佳的译文,对不佳的译文采用多策略增强方法,生成更多的伪数据,让数据集更有丰富性和多样性,进而训练出更好的翻译模型。

1 神经机器翻译

本文选择 Transformer 模型[1]进行实验,Transformer 模型是典型的端到端的深度学习模型架构,完全依赖于注意力机制对文本序列进行建模,注意力机制也能有效地解决长距离文本遗忘问题。Transformer 模型由编码器-解码器两部分组成,其中编码器和解码器都由多层相同的网络结构堆叠组成。

1.1 编码器

编码器中每一个网络层都包含 2 个子层,分别是多头注意力子层和前馈神经网络层。子层之间应用残差连接以及归一化处理。其核心运用了自注意力机制,通过计算得到句子在编码过程中每个位置的注意力权重,进而权重相加计算整个句子的隐含向量表示。具体计算过程是将词向量为 d 的输入分别映射到一组查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 的向量输出中,将这组向量输出进行点积操作,再通过 *Softmax* 计算权重,最后返回值的加权和,公式如下:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \quad (1)$$

为了使自注意力机制在处理序列数据时变得更加灵活和高效,提出了多头自注意力机制。其给予注意力层的输出包含不同子空间中的编码表示信息,让模型可以学习到更丰富的语义和结构特征。多头自注意力机制采用 h 个注意力头表示输入信息,将多头注意力的输出拼接乘以权重矩阵得到向量输出,公式如下:

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat(head_1,\cdots,head_h)\ W^O$$
$$head_{.i} = Attention(\boldsymbol{Q}W_i^Q,\boldsymbol{K}\ W_i^K,\boldsymbol{V}\ W_i^V) \quad (2)$$

其中, $W_i^Q, W_i^K, W_i^V, W_i^O \in \mathrm{R}^{d_{\mathrm{model}}\times d_k}$,$h$ 表示多头注意力机制中头的个数。

前馈神经网络层由 2 个线性变换组成,在 2 次线性变换中应用一次 *ReLU* 激活函数,推得的公式可写为:

$$FFN(x) = \max(0, x\boldsymbol{W}_1 + \boldsymbol{b}_1)\ \boldsymbol{W}_2 + \boldsymbol{b}_2 \quad (3)$$

其中, $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是学习矩阵, $\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 是随机偏置向量。

1.2 解码器

解码器的每一个网络层由 3 个子层组成。第 1 个子层为掩码多头注意力机制,第 2 个子层为编码器-解码器多头自注意力机制,第 3 个子层是前馈神经网络层,每 2 层之间经过残差以及归一化处理。对于各层设计,进行阐释解析如下。

第 1 层的掩码多头自注意力机制是多头自注意力机制的一种扩展形式,是在多头自注意力机制的基础上添加了掩码操作,假设序列 $Y(y_1,y_2,y_3,\cdots,y_n)$, 对 y_i 进行预测, 则掩码多头自注意力层只对 $(y_1,y_2,y_3,\cdots,y_{i-1})$ 进行注意力计算。

第 2 层的编码器-解码器多头自注意力机制的输入是掩码多头自注意力机制的输出和编码器的输出。将掩码多头自注意力的输出作为 $\boldsymbol{Q}$,编码器的输出作为 $\boldsymbol{K},\boldsymbol{V}$,进行注意力计算,使得解码过程中学

习源语言的信息。

第 3 层的前馈神经网络在计算上与编码器也类似，输入伪编码器-解码器多头注意力机制的输出。

2　多译文神经机器翻译的数据增强方法

该数据增强方法包含 2 个主要步骤。首先训练一个神经机器翻译模型，接着使用原始源端数据进行翻译，生成多个可能的译文，在对多译文筛选后，将利用 2 种策略生成伪双语数据，具体方法如下。

2.1　多译文生成策略

采用 beam-search 译文生成策略，该策略在每个时间步，生成多个可能的翻译结果，在以往的研究中，通常选择概率最大的结果作为翻译译文的输出。多译文生成见表 1。表 1 中，将希伯来句子“הים, מאד מסובך דבר להיות יכול הוא”翻译成多句英语句子。然而，概率略低的译文一部分并不是翻译有误，可能是同义词而导致的与参考译文不相符，但这并不表示翻译错误，如表 1 参考译文中的“ocean”在翻译译文 2 中被翻译为“sea”。因此，本文方法选取概率最大的前 2 个翻译译文。

表 1　多译文生成

Table 1　Multi-translation generation

内容	数据
源端数据	הים, מאד מסובך דבר להיות יכול הוא
参考译文	It can be a very complicated thing, the ocean.
翻译译文 1	It could be a very complicated thing for ocean.
翻译译文 2	It can be a complicated thing for sea.
翻译译文 3	It could be a complicated for illustration .

接着，为了有效地降低翻译质量较低的译文对训练翻译模型的影响，研究筛选出低质量的多译文来进行数据增强。本文设置固定的筛选指标，可由式(4)来描述：

$$P(ref, pre)=\frac{num[(t_{\text{ref}}^1, t_{\text{ref}}^2, \cdots, t_{\text{ref}}^n) \cap (t_{\text{pre}}^1, t_{\text{pre}}^2, \cdots, t_{\text{pre}}^n)]}{\text{num}(t_{\text{ref}}^1, t_{\text{ref}}^2, \cdots, t_{\text{ref}}^n)} \tag{4}$$

其中，$num[(t_{\text{ref}}^1, t_{\text{ref}}^2, \cdots, t_{\text{ref}}^n) \cap (t_{\text{pre}}^1, t_{\text{pre}}^2, \cdots, t_{\text{pre}}^n)]$ 是指参考译文与预测译文的重合单词个数，$num(t_{\text{ref}}^1, t_{\text{ref}}^2, \cdots, t_{\text{ref}}^n)$ 是参考译文单词个数。本文设置 $P(ref, pre) \leqslant 0.2$ 的句子筛选，进行数据增强。这种方法可以有效地降低翻译质量较低的译文对数据增强的影响，提高数据的质量和翻译的效果。

2.2　伪双语数据生成

伪双语数据是指在训练机器翻译模型时，通过对原始数据进行一些处理，生成一组看似双语的数据。

筛选生成多个译文后，为了有对应的源语，生成伪双语数据，本文采用 2 种策略，如图 1 所示。

2.2.1　生成伪数据策略 1

假设源数据 $S(s_1, s_2, \cdots, s_n)$，目标端数据 $T(t_1, t_2, \cdots, t_n)$，这里 n 表示语句数量。通过多译文生成策略选出 k 个($k < n$) 翻译不佳的源句和源句的翻译概率最高的 2 组译文，将其合并为一组伪双语数据。参见图 1 中的伪双语数据策略 1，其中 T_1' 与 T_2' 是翻译不佳的源句翻译出前 2 个概率最大的译文的集合，S_1 和 S_2 是译文对应的源语句的集合。

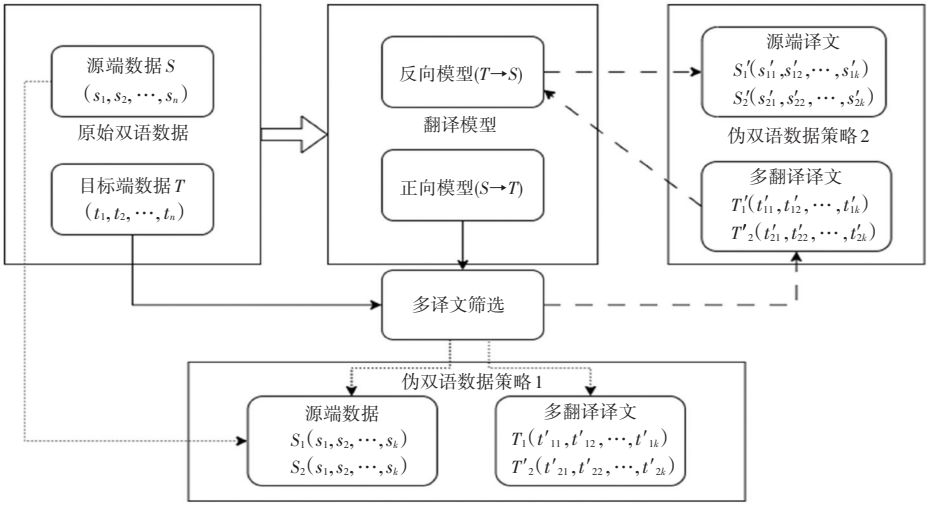

图 1　伪双语数据生成(2 种策略)

Fig. 1　Pseudo-bilingual data generation (two strategies)

2.2.2 生成伪数据策略 2

使用原始数据训练好的反向翻译模型,将生成的多翻译译文进行翻译,生成伪双语数据,参见图 1 中的伪双语数据策略 2,其中 S_1' 和 S_2' 是反向翻译后生成的源语句的集合, T_1' 和 T_2' 与策略 1 相同。

3 实验分析

3.1 数据预处理与模型设置

3.1.1 数据预处理

为验证本文数据增强方法在稀缺资源下的有效性,对翻译任务进行了多次实验。实验采用公开稀缺资源数据集 IWSLT14 希伯来语-英语(He-en),IWSLT15 越南语-英语(Vi-en)。其中,IWSLT14 希伯来语-英语训练任务,使用了 181 k 的平行句对作为训练集,tst2013 设置为校验集,tst2014 设置为测试集。对于 IWSLT15 越南语-英语的训练任务使用 tst2012 作为验证集,tst2013 作为测试集。

为了有效地训练实验模型,本文对于所有语料库数据进行数据预处理。所有语料都进行了规范化(normalize)、符号化(tokenize)以及 BPE(Byte Pair Encoding)[12]子词切分等处理。本文采用双语句对的联合词表进行 10 k 的 BPE。表 2 统计了各语言训练集、校验集和测试集。

表 2 语料大小

Table 2 Corpus size

语料	统计内容		
	训练集	验证集	测试集
He-en	181 k	1 199	1 305
Vi-en	133 k	1 553	1 268

3.1.2 模型设置

本文使用 Transformer_small 模型,Adam[13]优化器,其中参数 $\beta_1 = 0.9$,$\beta_2 = 0.98$。学习率调度器为 inverse-sqrt,权重衰减为 0.000 1,wormup 步数为 4 000。使用交叉熵损失函数作为损失函数,并设置 *dropout* 率为 0.3。为了适应不同长度的句子,按照句子长度划分不同的批次,并设置每块 GPU 的最大输入输出为 4 096 个词。实验使用 GTX3080、GTX3090TI 训练模型,翻译性能使用 *Bleu*[14] 评测。

3.2 实验结果和分析

为了验证本研究提出的多译文数据增强方法的有效性,分别在 Transformer 基准模型上进行比较实验,本实验对多译文取了 2 个样本,分别是 TOP1 和 TOP2,其中 TOP2 包括 TOP1 的译文。

3.2.1 多译文伪双语生成策略 1

本文提出的基于多译文双语生成策略 1 的数据增强方法在稀缺资源翻译任务结果见表 3。分析可知,与基线相比,TOP1 译文和 TOP2 译文分别在 He-en,Vi-en 提升了 0.27,0.24,0.47,0.31 个 *Bleu* 值。总结得出:

(1)保持源端不变的情况下,筛选出的质量不佳的多译文进行数据增强时,可以使翻译模型效果提升,同时 TOP2 的译文数据增强使得翻译模型效果提升更加显著。

(2)进一步分析,本增强方法扩充了原始数据集,同时伪数据是由翻译模型翻译的多译文生成,故伪数据具有一定的噪音,提升了模型的鲁棒性。

表 3 伪双语生成策略 1 实验结果

Table 3 Experimental results of pseudo-bilingual generation strategy 1

方法	语料	
	He-en	Vi-en
基线	34.00	31.28
TOP1+策略 1	34.27	31.52
TOP2+策略 1	34.47	31.59

3.2.2 多译文伪双语生成策略 2

表 4 展示了多译文生成后利用反向翻译生成对应译文的双语伪数据增强方法。与基线相比,TOP1 在 2 个数据集上有更明显的提升,在 He-en 数据集上提升了 0.64 个 *Bleu* 值,在 Vi-en 数据集上提升了 0.35 个 *Bleu* 值。TOP2 在实验中与 TOP1 相比,几乎持平、甚至下降。总结得出:TOP1 的伪双语数据质量更佳,对翻译模型提升的效果更明显。

表 4 伪双语生成策略 2 实验结果

Table 4 Experimental results of pseudo-bilingual generation strategy 2

方法	语料	
	He-en	Vi-en
基线	34.00	31.28
TOP1+策略 2	34.64	31.60
TOP2+策略 2	34.50	31.62

3.2.3 多译文伪双语与原始双语数据增强对照

为了验证本方法生成更加丰富的伪双语数据,增加了混合原始双语(Src-Tar-DA)实验。具体地,对翻译后的译文与参考译文比较,在筛选译文后、本

实验不再对翻译后的译文数据增强，而是将翻译欠佳的参考原文和参考译文抽取出作为伪双语数据。实验结果见表 5。本文提出的多译文数据增强方法比 Src-Tar-DA 方法对模型提升的效果更显著，进一步得出本方法的有效性。

表 5　Src-Tar-DA 比较实验结果

Table 5　Src-Tar-DA comparative experimental results

方法	语料	
	He-en	Vi-en
基线	34.00	31.28
TOP1+策略 1	34.27	31.52
TOP2+策略 2	34.64	31.60
Src-Tar-DA	34.06	31.39

3.2.4　与其他数据增强方法比较

为进一步验证本文提出的数据增强方法的有效性，本文与其他现有的数据增强方法进行比较，结果见表 6。为了实验的公平性，本文实验对数据集的设置与 Gao 等学者[15]文献中的保持一致。很明显，本文提出的数据增强方法优于其他数据增强方法，可以证明本文方法的鲁棒性。

表 6　其他数据增强方法比较实验结果

Table 6　Comparison of experimental results with other data enhancement methods

方法	He-en
Base	34.00
Swap	34.25
Blank	34.37
Drop	34.29
TOP1+策略 2	34.64

3.2.5　训练损失与 *Bleu* 的比较

为了验证基于本文数据增强方法的收敛性能，本文计算了 IWSLT14 越南语-英语翻译任务在不同时期的训练损失和 *Bleu* 分数。IWALT14 越南语-英语基线与 TOP1-策略 2 训练损失比较结果曲线如图 2 所示，IWALT14 越南语-英语基线与 TOP1-策略 2*Bleu* 值比较结果曲线如图 3 所示。由图 2、图 3 可以看到，随着训练轮数（*epoch*）的增加，*Bleu* 分数提升，损失值下降，最终逐渐平稳。与基线相比，本文的方法可以更快地收敛。以此推断，伪平行语料丰富训练集的多样性，神经机器翻译模型表现更好。

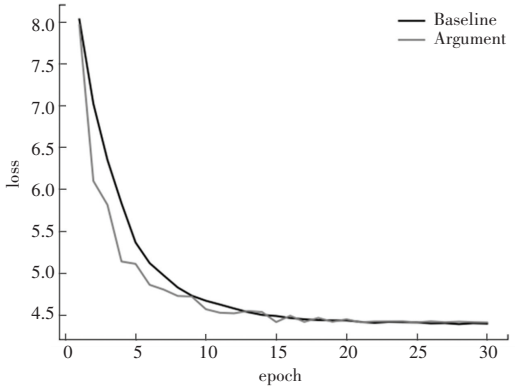

图 2　IWALT14 越南语-英语基线与 TOP1-策略 2 训练损失比较

Fig. 2　Training loss comparison between Vietnamese - English baseline and TOP1-strategy 2 of IWALT14

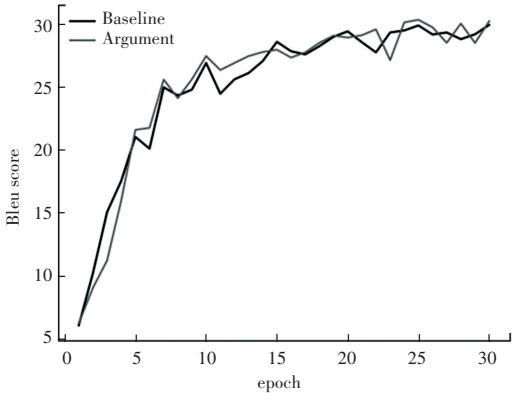

图 3　IWALT14 越南语-英语基线与 TOP1-策略 2 *Bleu* 值比较

Fig. 3　Comparison of *Bleu* between Vietnamese-English baseline and TOP1-strategy 2 of IWALT14

4　结束语

本文提出了一种有效的神经机器翻译数据增强方法。融合多译文与机器翻译任务生成伪双语数据。通过在稀缺数据集上的翻译实验，可以看出，相对于基线系统，该方法显著地提升了模型的翻译性能。进一步验证，所提出的方法比其他现有的数据增强部分方法效果更好。在未来的工作中，继续探索 K 个（$K > 2$）译文样本生成伪数据对翻译模型的影响，或者研究文中提出方法在自然语言处理的其他生成任务中的应用。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017, 30: 5998-6008.

[2] PONCELAS A, POPOVIC M, SHTERIONOV D, et al. Combining SMT and NMT back-translated data for efficient NMT[J]. arXiv

preprint arXiv:1909.03750, 2019.
[3] 李响,刘洋,陈伟,等. 利用单语数据改进神经机器翻译压缩模型的翻译质量[J]. 中文信息学报,2019,33(7): 1-10.
[4] JASON W, ZOU Kai. EDA: Easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.
[5] WANG Xinyi, PHAM H, DAI Zihang, et al. SwitchOut: An efficient data augmentation algorithm for neural machine translation[J]. arXiv preprint arXiv:1808.07512, 2018.
[6] FADAEE M, BISAZZA A, MONZ C. Data augmentation for low-resource neural machine translation[J]. arXiv preprint arXiv: 1705.00440, 2017.
[7] MUHAMMAD M, LIU Yang, LUAN Huanbo, et al. Improving data augmentation for low-resource NMT guided by POS-tagging and paraphrase embedding[J]. Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(6): 1-21.
[8] WU Xing, LV Shangwen, ZANG Liangjun, et al. Conditional bert contextual augmentation[C]// 19th International Conference on Computational Science (ICCS 2019). Faro, Portugal: Springer International Publishing, 2019: 84-95.
[9] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[J]. arXiv preprint arXiv:1511.06709, 2015.
[10] ZHANG Yi, GE Tao, SUN Xu. Parallel data augmentation for formality style transfer[J]. arXiv preprint arXiv: 2005.07522, 2020.
[11] HOANG V C D, KOEHN P, HAFFARI G, et al. Iterative back-translation for neural machine translation[C]//Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. Melbourne, Australia: ACL, 2018: 18-24.
[12] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[J]. arXiv preprint arXiv:1508.07909, 2015.
[13] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[14] 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. Philadelphia, USA: ACL, 2002: 311-318.
[15] GAO Fei, ZHU Jinhua, WU Lijun, et al. Soft contextual data augmentation for neural machine translation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 5539-5544.