文章编号：2095-2163(2024)03-0067-09 中图分类号：TP391 文献标志码：A

基于词典的低资源神经机器翻译数据增强方法

张宝兴
(上海理工大学 光电信息与计算机工程学院，上海 200093)

摘 要：数据增强是提升低资源语种上神经机器翻译性能的有效手段，传统的回译方法能够有效利用目标语言的单语数据对模型进行训练，但是由于回译模型的质量与可用的平行语料库大小有关，导致在低资源场景下生成的伪平行语料质量较差。针对以上问题，本文提出了一种基于词典的低资源神经机器翻译数据增强方法，首先从平行语料中抽取词典；其次，在平行语料和目标语言的单语语料中选取合适的模版句子，并对其中的单词进行替换，从而生成伪平行语料以辅助神经机器翻译模型的训练。在公开数据集上的实验证明：使用该数据增强方法处理后的数据集，能够使基线翻译模型获得 3.71-6.42 的 *BLEU* 值提升。

关键词：低资源语种；神经机器翻译；数据增强

Dictionary based data augmentation method for low-resource NMT

ZHANG Baoxing
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Data augmentation is an effective approach to improve the performance of neural machine translation on low-resource languages. The traditional back-translation method can effectively use the monolingual data of the target language to train the model, but because the quality of the back-translation model is related to the size of the available parallel corpus, the quality of the pseudo-parallel corpus generated in the low-resource scenarios is poor. Aiming at the aboved problems, this paper proposes the dictionary-based low-resource neural machine translation data enhancement method. The method firstly extracts the dictionary from the parallel corpus, and then selects the appropriate template sentence from the parallel corpus and the monolingual corpus of the target language and replace the word(s) in the selected sentence to generate a pseudo-parallel corpus to assist the training of the neural machine translation model. Experiments carried out on public datasets prove that the dataset augmented by the proposed method can improve the baseline translation model by 3.71—6.42 in *BLEU* value.

Key words: low-resource languages; neural machine translation; data augmentation

0 引 言

神经机器翻译(NMT)模型是一种端到端模型，包含一个编码器和一个解码器，编码器的作用是将输入的源语句转化为向量表示；解码器则利用生成的 LSTM(Long Short-Term Memory)隐藏状态和注意力机制，生成目标语言的翻译结果。一般而言，机器翻译系统都需要大量的训练数据，使得最终训练出的翻译模型在实际使用中取得较好的泛化效果。训练数据一般指源语言与目标语言之间的平行语料。在低资源场景下执行机器翻译任务时，一个最大的障碍是平行语料资源匮乏问题。如何在低资源场景下利用好有限的平行语料资源，及相对丰富的单语语料资源，以提升神经机器翻译模型的训练效率及最终的模型质量成为了一个重要的研究方向。数据增强(Data Augmentation)是一类利用现有数据创造额外数据或从不同数据源添加数据以辅助深度学习模型进行训练的方法。对于机器翻译研究而言，数据增强被广泛用于生成伪平行语料以训练神经机器翻译模型。在该领域的研究中，Sennrich 等学者[1]提出的回译(Back Translation)是其中最有效的方法之一，基于平行语料数据训练一个反向的翻译模型，将目标语言的单语语料，翻译成源语言，从而扩充原有的平行语料数据。Currey 等学者[2]提出复制目标

基金项目：国家自然科学基金(61772342)。

作者简介：张宝兴(1996-)，男，硕士研究生，主要研究方向：自然语言处理。Email：zhangbaoxing96@163.com

收稿日期：2023-03-05

语言单语语料的方法(Copy),直接将目标语言的单语数据复制作为源语言语句,组成平行语料并输入模型进行训练,目标是捕捉不同语言之间的相同性质。以上提到的 2 种方法都存在不足之处:使用回译模型生成的源语句,其质量与平行语料的数量呈正相关,即在平行语料充足的情况下,才能保证其训练的反向翻译模型的质量,对于模型而言能够学习的词汇范围也仅限于平行语料中出现过的词汇,这将导致出现 OOV(Out-Of-Vocabulary)现象的可能性,使得模型在实际使用中的效果不好;而 Copy 方法也只关注了目标语言的词汇。

本文提出了一种基于词典的低资源神经机器翻译数据增强方法(Low-Resource Dictionary based Data Augmentation, LDDA),该方法可对平行语料及单语语料进行数据增强,从而显著提升基线神经机器翻译模型的性能。通过实验证明了根据从平行语料中抽取的字典,对平行语料及单语语料中被选中的词汇/词组执行替换以自动生成伪平行语料是一种高效并可行的方法。

1 相关工作

神经机器翻译的数据增强技术可以分为 3 个类别:基于单词或词组替换的数据增强、基于回译(back translation)方法的数据增强以及相似语料数据挖掘。

本文基于单词或词组替换的数据增强技术,从现有的平行语料或单语语料中选取一部分候选句子,通过替换这些候选句中的部分单词或词组,从而生成句子。一种方式是针对单语语料使用一个双语词典,并用目标语言的单词替换候选句子中所有的单词,以生成翻译的结果。Nag 等学者[3]尝试将句子中所有的单词都进行替换。Peng 等学者[4]则只替换句子中较为罕见的单词。Zhang 等学者[5]尝试使用双语词典实现数据增强,但研究主要关注罕见词汇,其方法依赖基于词组的统计翻译模型。Arthur 等学者[6]在研究中引入了离散翻译词典(discrete translation lexicons)的概念,其目的是为了影响模型解码器的概率分布。Fadaee 等学者[7]尝试使用另一种解决方式,即将目标语言中的高频词汇替换成罕见词汇,随后调整对应的源语句词汇。Waldendorf 等学者[8]在采用联合学习方法的基础上,引入单语数据和双语词典,改进双语词嵌入,并结合逐词反向翻译方法,提升平行语料中没有出现过或很少出现过的词的翻译质量。

2016 年,Sennrich 等学者[1]提出回译方法,成为了利用目标语言侧单语语料进行数据增强的一种主流方法。Burlot 等学者[9]研究证明,该方法中的反向翻译模型会直接影响翻译质量—使用一个较弱的回译模型进行增强的数据,只会对模型效果产生细微的提升作用。本文提出的方法直接使用从给出的平行语料中提取出的双语词典解决了这一问题。Artetxe 等学者[10]和 Hoang 等学者[11]尝试使用迭代式的回译方法,源语言和目标语言的单语数据分别使用源语言→目标语言和目标语言→源语言的翻译模型进行翻译,这个过程持续迭代进行,同样的一组句子会被回译多次,直到回译方法产生的数据不再对目标翻译模型产生提升作用为止 。但是在这些方法中,2 个方向上的翻译模型是分别独立进行训练的。为了解决这一问题,Zheng 等学者[12]提出了一个可以将这 2 个方向上的模型进行联合翻译的模型。另外,在回译方法中,简单将所有的单语数据进行回译是不能保证取得最优结果的,Edunov 等学者[13]提出了源-生成数据比例指数的概念,并证明了该指数对于回译的表现能够起到衡量作用。所以,在使用回译方法时,需要谨慎控制生成的伪平行语料与原有的平行语料之间的比例。Dou 等学者[14]和 Fadaee 等学者[15]在这种思路的引导下,从单语语料中选择了最合适的一个子集进行了回译操作,取得了不错的实验效果。

相似语料指的是描述同一主题的文字,虽然互相之间并非直接翻译的结果,但是可能包含了互为翻译的片段,如使用不同语言描述同一事件的维基百科和新闻语料。从相似语料中提取出的平行语句一直都被认为是为机器翻译提供生成语句的一种很好的来源。早期的研究主要使用神经机器翻译的编码器-解码器结构生成多语言的句子嵌入表示,Yang 等学者[16]尝试使用双语多编码器结构。Schwenk[17]尝试使用共享多语种编码器-解码器结构。Artetxe 等学者[18]通过将多种语言联合嵌入到一个共享空间,并开源了第一个成功探索大型多语言句子表示的开源 NLP( )工具,该工具支持 93 种语言并成为了后续大规模平行语料抽取任务的基础。

其他数据增强方法包括由 Currey 等学者[2]提出的将目标语言数据直接复制作为源语言语句中使用的 Copy 方法。Gupta 等学者[19]通过对平行语料进行句法分析,生成源语言和目标语言之间的短语片段,并以此扩充训练语料库。与前述研究不同的

是，本文从平行语料中抽取出的是双语词典。

2　模型方法

本文提出的 LDDA 数据增强方法主要包含了语句对齐、语句嵌入计算、词嵌入计算、词匹配及词替换等步骤，使用该方法对平行语料进行数据增强处理的流程示意图如图 1 所示。首先从平行语料中抽取出语料词典，从平行语料及目标语言的单语语料中挑选合适的模版语句，并将其中的单词进行替换，从而生成伪平行语料，供神经机器翻译模型进行训练。

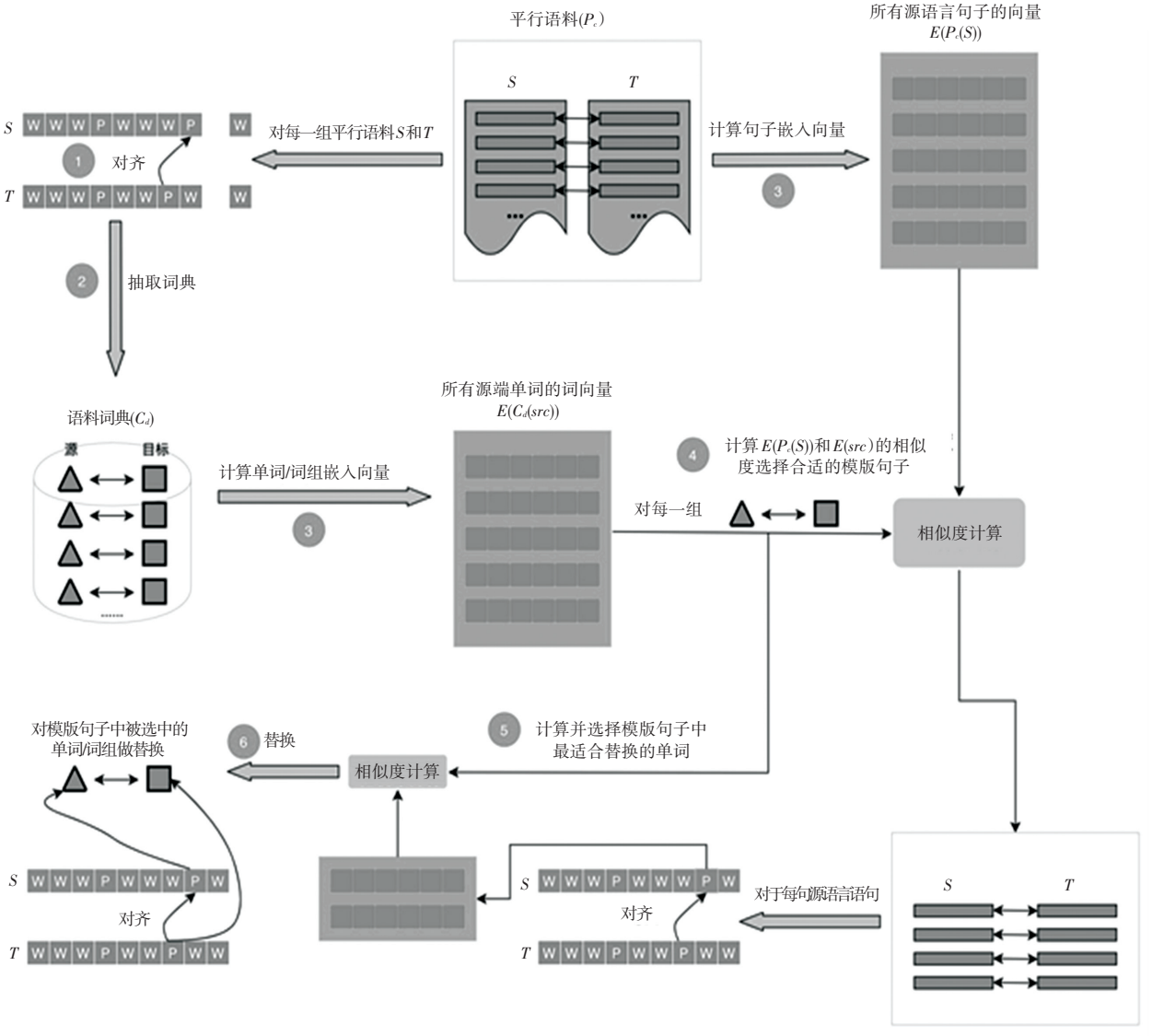

图 1　使用本文提出的方法处理平行语料的流程

Fig. 1　Proposed model dealing with parallel corpus

2.1　语料对齐和词典抽取

首先，基于 FastAlign 工具[20]对给定的有限平行语料训练一个词级别的语句对齐模型，语句对齐模型在随后的词典抽取以及最后的单词/词组替换步骤中都起到了很关键的作用。基于对齐的结果，在有限的平行语料上抽取出一个语料词典（C_d）。

2.2　单词/词组和句子嵌入表示

使用 bert-as-service 项目[21]计算语料词典和平行语料中句子的嵌入表示。该项目基于预训练的多语种大小写敏感的 bert-base 模型[22]。源语言词组的嵌入信息通过平行语料字典获得，记作 $E(C_d(src))$；源语句的嵌入信息通过平行语料获得，记作

$E(P_c(src))$。本文中只针对语料词典中的源语言单词/词组(在图 1 中标记为 *src*) 以及平行语料中的源语言句子计算嵌入表示并使用, 因为如果使用目标语言的单词/词组及平行语料中的目标语言句子进行计算,其生成的平行语料将与使用源语言的单词/词组及平行语料中源语言句子生成的伪平行语料非常类似,这对于翻译模型的训练而言将不会产生任何的增益效果。

2.3 词组-句子匹配

词组-句子匹配是为了从有限的平行语料中挑选出合适的句子作为模板,进而基于语料词典进行替换。对于语料词典中的每一个源语言单词嵌入,本文尝试根据句子嵌入与源语言单词嵌入的相似度计算寻找前 N 个源语言句子 $((S_k)_{k=1}^{N})$,由此推得:

$$(S^k)_{k=1}^{N} = k - \underset{j}{\operatorname{argmin}} \frac{E(src) * E(S^j)}{\| E(src) \| * \| E(S^j) \|} \quad (1)$$

考虑到数据量比较大,本文使用 Faiss 方法进行高效的相似度搜索。Faiss 是一种通过在密集向量的聚类上开发索引来在大型数据集中进行相似性搜索的库方法,利用余弦相似度(内积)被用于聚类和搜索过程[23]。以上步骤将生成一系列模板句子 $(S_k)_{k=1}^{N}$ 以供下一步操作。

2.4 单词/短语匹配和替换

每个模板句子的短语在词元(tokenization)化后使用 TextBlob 提取,并将这些模版句子中短语的嵌入与源短语的嵌入进行比较。算法 1 中的 *top_sim* 函数从语料库字典中提取与匹配单词/短语具有最大余弦相似度的单词/短语 $(P_{\max}(S^k))$, 可由式(2)进行计算:

$$P_{\max}(S^k) = \underset{p}{\operatorname{argmin}} \frac{E(src) * E(P(S^k))}{\| E(src) \| * \| E(P(S^k)) \|} \quad (2)$$

本文仅关注名词和动词,使用在之前词典抽取环节中训练的对齐模型,定位到与源语言的模板语句中的候选单词/短语 $P_{\max}(S^k)$ 相对应的目标语句中相应的单词/词组 $P_{\max}(T^k)$。通过替换模板语句 $\{S^k, T^k\}$ 中的候选单词/词组对 $\{P_{\max}(S^k), P_{\max}(T^k)\}$ 即可生成伪平行语料对,替换过程需要使用之前提取出的语料词典,并通过语料词典找到单词/词组的对应翻译 $\{src^i, tgt^i\}$。通过重复这个过程,遍历所有的词典词条以生成伪平行语料。

算法 1 桥接迭代回译训练算法

输入 $P_c = \{(S^j, T^j)\}_{j=1}^{n}$// 平等语料

输出 $G_c = \{(G_{src}^k, G_{tgt}^k)\}_{k=1}^{o}$// 生成的伪平行语料

```
G_c ← Ø
for (S^j, T^j) ∈ {(S^j, T^j)}_{j=1}^{n} do
    align ((S^j, T^j))
    C_d ← [S_k^j, T_k^j]
end for
//计算词典中的单词和句子的嵌入
E(C_d(src)) = E(src^i)_{i=1}^{m}
E(P_d(S)) = E(S^j)_{j=1}^{n}
for E(src) ∈ E(src^i)_{i=1}^{m} do
    //选出 TopN 个与当前单词最匹配的模版句子
    (S^k)_{k=1}^{N} → faiss(E(src), E(S^j)_{j=1}^{n})
    //遍历 TopN 个模版句子
    for S^k ∈ (S^k)_{k=1}^{N} do
        //抽取出每个句子中的单词/词组
        P(S^k) ← phrase_extract(S^k)
        //计算出目标句子中与当前单词相似度最高的单词/词组
        P_max(S^k) ← top_sim(E(src), E(P(S^k)))
        //根据对齐结果,定位到要替换的单词/词组
        P_max(T^k) ← algin(P_max(S^k))
        //完成最终的替换
        G_c(S) ← sub(S^k, P_max(S^k), src)
        G_c(T) ← sub(T^k, P_max(T^k), tgt)
        G_c = G_c ∪ {G_c(S), G_c(T)}
    end for
end for
return G_c
```

2.5 单语语料处理

本文还探索了如何利用单语数据生成伪平行语料。通过复用在平行语料上提取出的语料库词典对单语语料句子中的单词进行替换,受 Currey 等学者[2]在低资源神经机器翻译中使用目标语言单语语料创建伪平行语料的启发,将替换单词后的目标语言语句直接复制作为源语句组成平行语料。使用 LDDA 方法对单语语料进行数据增强的流程如图 2 所示。图 2 中,虚线包围的部分表示该过程已在先前对平行语料进行数据增强生成伪平行语料的步骤中完成。

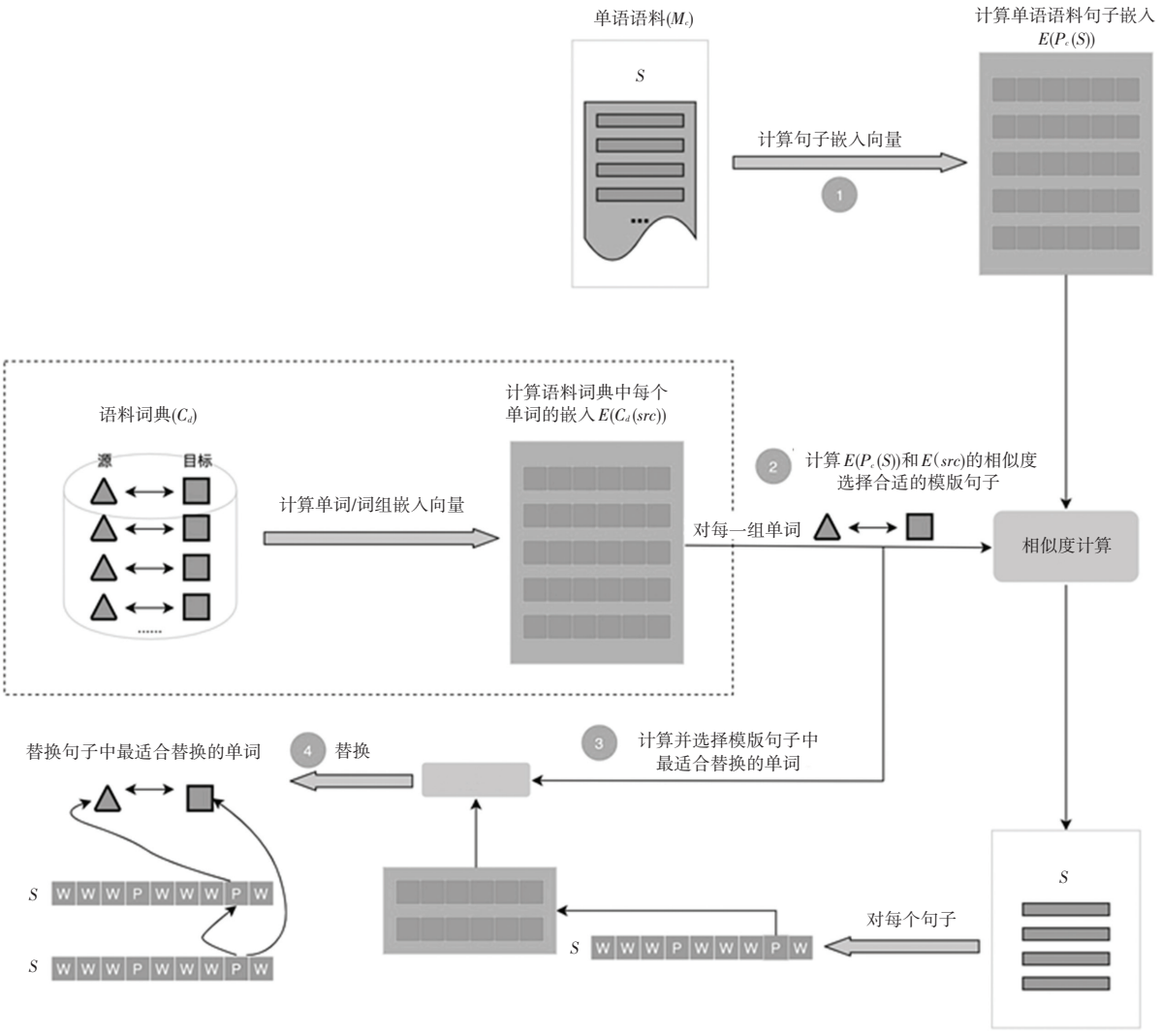

图 2　使用本文提出的方法处理单语语料的流程

Fig. 2　Proposed model dealing with monolingual corpus

3　实验

3.1　实验设置

本文使用 wmt18 数据集在德语-英语(de-en)和法语-英语(fr-en)翻译任务上开展实验,从给定的训练集中随机采样生成数据子集以模拟低资源场景。实验选用的基线神经机器翻译模型由基于 transformer 架构的编码器和解码器组成,模型的超参数见表 1。使用 Moses 项目对生成的数据及有限平行语料资源进行清理和预处理,将语料中的标点符号规范化为标准形式,并将句子中的单词进行词元化,从而分解为可处理单元;训练大小写模型以调整句子中单词的大小写;使用字节对编码(Byte-Pair Encoding)方法处理语料中的罕见词汇。

表 1　基线神经机器翻译模型的超参数

Table 1　Hyperparameters for baseline NMT model

超参数	数值
编码层数量	6
解码层数量	6
词嵌入维度	2 048
注意力头个数	8
前馈神经网络隐藏单元个数	512
初始学习率	0.006

首先,使用有限的平行语料数据对基线神经机器翻译模型进行训练;使用不同的数据增强方法对数据集执行数据增强;使用生成的增强数据集对基线神经机器翻译模型再进行一次训练;在测试集上测试、对比前后 2 次训练出的模型性能,从而证明数

据增强方法的有效性。将 LDDA 在平行语料上执行数据增强生成的伪平行语料、记作 LDDA-para，LDDA 在单语语料上执行数据增强生成的伪平行语料、记作 LDDA-mono。

本文共进行了 4 项实验，以证明所提出的 LDDA 数据增强的有效性。这里给出阐释分述如下。

(1)不同数据增强方法间的性能对比。为了与基于词典的数据增强的相关工作进行比较，实验对比了 Nag 等学者[3]提出的逐字替换伪平行语料生成方法(Word-on-Word data augmentation, WoW)，该方法使用公开可用的双语词典进行逐字翻译以生成伪平行语料。另外，还对比了传统的数据增强方法、即回译方法及 Copy 方法。

(2)平行语料增强数据规模对模型性能的影响。本项实验主要研究使用 LDDA 数据增强框架生成不同规模的平行语料增强数据，对翻译模型的性能影响。

(3)单语语料增强数据对模型性能的影响。本项实验主要研究在使用 LDDA 数据增强框架生成平行语料增强数据的基础上，增加生成的单语语料增强数据，对翻译模型的性能影响。

(4)LDDA 与其他数据增强方法的互补性。本项实验主要研究 LDDA 数据增强框架与其他数据增强方法(如 Copy 和回译)的互补性，结合不同的数据增强方法生成的增强数据在基线神经机器翻译模型上开展实验，对比模型性能。

实验采用 *BLEU* (Bilingual Evaluation Understudy)值作为评价指标，评估模型的翻译性能。

3.2 实验结果和分析

3.2.1 翻译性能

本将 LDDA 方法与 3 个基线模型进行对比：

(1)Parallel：仅 10 k 句平行语料 。

(2)BT：10 k 句平行语料+10 k 句使用回译方法增强的语句对。

(3)Copy：10 k 句平行语料+10 k 句 Copy 方法复制的目标语言语料。

(4)WoW：10 k 句平行语料+10 k 句 WoW 方法增强的语句对。

实验将 LDDA 方法在以下 6 种原始和增强数据集上进行训练的性能对比。这里，字母 w 表示单位：万。详细内容阐述如下。

(1)1 w Parallel：1 w 句平行语料。

(2)1 w Parallel+1 w BT：1 w 句平行语料+1 w 句使用回译方法增强的语句对。

(3)1 w Parallel+1 w copy：1 w 句平行语料+1 w 句 Copy 方法复制的目标语言语料。

(4)1 w Parallel+1 w WoW：1 w 句平行语料+1 w 句 WoW 方法增强的语句对。

(5)1 w Parallel+1 w LDDA-para：1 w 句平行语料+1w 句使用 LDDA 方法对平行语料增强生成的伪平行语料。

(6)2 w Parallel：2 w 句平行语料。

实验以 1 w 句平行语料组成的数据集作为基准，并使用不同数据增强方法生成伪平行语料，通过基线翻译模型在原始数据集及不同方法增强的数据集上进行训练，对比模型性能说明数据增强方法的有效性，基线模型在使用不同数据增强方法获取到的数据集上的 *BLEU* 值见表 2。从表 2 可知，使用 BT 方法增强后的数据集训练的基线翻译模型其 *BLEU* 值优于仅使用平行语料(Parallel)进行训练的模型，这表明即使数据增强方法生成的伪平行语料质量较差，也能够在一定程度上提升翻译模型的性能；Copy 方法优于 BT 方法，BT 方法的数据增强效果较差可归因于其生成的伪平行语料源语句质量低于 Copy 方法；将使用 LDDA 方法对平行语料进行数据增强生成的伪平行语料记作 LDDA-para，在 en→de 和 en→fr 这两个翻译方向上，模型性能优于使用其他方法增强的数据集训练的模型；LDDA 方法在 en→de 和 en→fr 这 2 个翻译方向上，相较 Copy 方法在 *BLEU* 指标上分别提升了 1.91 和 2.04，相较 WoW 方法则分别提升了 1.02 和 1.23。表 2 中最后一行的数据为基线翻译模型在 2 w 句平行语料数据集上训练的性能指标，在原有 1 w 句平行语料数据集上进行数据增强可实现的性能提升上限。

表 2 基线模型在使用不同数据增强方法获取到的数据集上的 *BLEU* 值

Table 2 *BLEU* values of baseline NMT model on the data-augmented datasets

实验组	en→de	en→fr
1 w Parallel	9.34	9.65
1 w Parallel+1 w BT	10.77	11.51
1 w Parallel+1 w Copy	11.14	11.53
1 w Parallel+1 w WoW	12.03	12.34
1 w Parallel+1 w LDDA-para	**13.05**	**13.57**
2 w Parallel	13.47	13.68

3.2.2　增强数据的数量对模型的影响

实验通过对比在平行语料上进行数据增强，生成不同规模的伪平行语料加入训练对基线翻译模型性能的影响。基于原始平行语料与平行语料增强生成的伪平行语料之间的不同比例，共进行了 5 组实验，这里字母 w 表示单位：万。对此可做具体表述如下。

(1)1 w Parallel：1 w 句平行语料。

(2)1 w Parallel+1 w LDDA-para：1 w 句平行语料+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料。

(3)1w Parallel+2 w LDDA-para：1 w 句平行语料+2 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料。

(4)1 w Parallel+4 w LDDA-para：1 w 句平行语料+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料。

(5)2 w Parallel：2 w 句平行语料。

基线模型在使用不同规模 LDDA 增强数据的数据集上的 *BLEU* 值见表 3。从表 3 中可知，随着 LDDA 数据增强的平行语料数量的增加，基线翻译模型的性能有相应提升，其中 en→de 方向上的 *BLEU* 值最高提升了 6.42，en→fr 数据集上的 *BLEU* 则最高提升了 6.18。表 3 中的最后一行数据为基线翻译模型在 2 w 句平行语料数据集上进行训练的性能指标，可以观察到 1w Parallel+2w LDDA-para 数据集在基线模型上训练后达到的效果已经超过了该指标，证明了 LDDA 方法的优越性。

表 3　基线模型在使用不同规模 LDDA 增强数据的数据集上的 *BLEU* 值

Table 3　*BLEU* values of baseline NMT model on different amount of LDDA-augmented dataset

实验组	en→de	en→fr
1 w Parallel	9.34	9.65
1 w Parallel+1 w LDDA-para	13.05(+3.71)	13.57(+3.92)
1 w Parallel+2 w LDDA-para	14.55(+5.15)	14.86(+5.21)
1 w Parallel+4 w LDDA-para	15.76(+6.42)	15.83(+6.18)
2 w Parallel	13.47	13.68

3.2.3　单语增强数据对模型的影响

实验主要探究单语语料数据增强方法生成的伪平行语料对基线翻译模型训练效果的影响。将应用 LDDA 方法对单语数据增强生成的伪平行语料记作 LDDA-mono，基于原始平行语料、平行语料增强生成的伪平行语料(LDDA-para)及单语语料增强生成的伪平行语料(LDDA-mono)之间的不同比例，共进行了 4 组实验，这里字母 w 表示单位：万。研究论述详见如下。

(1)1 w Parallel：1 w 句平行语料。

(2)1 w Parallel+1 w LDDA-para+1 w LDDA-mono：1 w 句平行语料+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料+1 w 句使用 LDDA 方法对单语语料增强生成的伪平行语料。

(3)1 w Parallel+1 w LDDA-para+2 w LDDA-mono：1 w 句平行语料+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料+2 w 句使用 LDDA 方法对单语语料增强生成的伪平行语料。

(4)2 w Parallel：2 w 句平行语料。

基线模型在使用引入平行语料数据和单语数据 LDDA 增强数据的数据集上的 *BLEU* 值见表 4。从表 4 可知，当原始数据集中同时加入 LDDA-mono 与 LDDA-para 伪平行语料时，基线翻译模型的性能进一步提升。表 4 中最后一行数据仍为基线翻译模型在 2 w句平行语料数据集上进行训练的性能指标，当数据比例为 1∶1∶1(1 w Parallel+1 w LDDA-para+1 w LDDA-mono)时，基线模型的性能已经超越了该指标，说明了 LDDA 数据增强方法对于改善 OOV(Out-Of-Vocabulary)问题的有效性，使用 LDDA 方法分别对平行语料及单语语料数据进行增强可以让模型更好地学习语料中的罕见词汇，从而改善模型的泛化性能。

表 4　基线模型在使用引入平行语料数据和单语数据 LDDA 增强数据的数据集上的 *BLEU* 值

Table 4　*BLEU* values of baseline NMT model on LDDA-augmented dataset introducing both parallel and monolingual data

实验组	en→de	en→fr
1 w Parallel	9.34	9.65
1 w Parallel+1w LDDA-para+ 1 w LDDA-mono	13.05(+3.71)	13.57(+3.92)
1 w Parallel+1w LDDA-para+ 2 w LDDA-mono	14.55(+5.15)	14.86(+5.21)
2 w Parallel	13.47	13.68

3.2.4　将 LDDA 与其他数据增强方法相结合

实验将 LDDA 与其他数据增强方法(如 Copy 和 BT)相结合，验证其与其他数据增强方法的互补性，

共进行了 7 组实验,这里字母 w 表示单位:万。文中拟做概述如下。

(1)1 w Parallel:1 w 句平行语料。

(2)1 w Parallel+1 w BT:1 w 句平行语料+1 w 句使用回译方法增强的语句对。

(3)1 w Parallel+1 w Copy:1 w 句平行语料+1 w 句 Copy 方法复制的目标语言语料。

(4)1 w Parallel+1 w BT+2 w WoW:1 w 句平行语料+1 w 句使用回译方法增强的语句对+2 w 句 WoW 方法增强的语句对。

(5)1 w Parallel+1 w BT+1 w LDDA-para+1 w LDDA-mono:1 w 句平行语料+1 w 句使用回译方法增强的语句对+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料+1 w 句使用 LDDA 方法对单语语料增强生成的伪平行语料。

(6)1 w Parallel+1 w Copy+2 w WoW:1 w 句平行语料+1 w 句 Copy 方法复制的目标语言语料+2 w 句 WoW 方法增强的语句对。

(7)1 w Parallel+1 w Copy+1 w LDDA-para+1 w LDDA-mono:1 w 句平行语料+1 w 句 Copy 方法复制的目标语言语料+1 w 句使用 LDDA 方法对平行语料增强生成的伪平行语料+1 w 句使用 LDDA 方法对单语语料增强生成的伪平行语料。

Currey 等学者[2]已经通过实验验证了 Copy 方法与 BT 方法相结合能够提升数据增强效果。通过对比 WoW 及 LDDA 与 BT 和 Copy 方法相结合对数据集进行增强时基线翻译模型的性能,说明 LDDA 方法能够更好地与其他方法结合对数据集进行增强,基线模型在结合不同数据增强方法获取到的数据集上的 *BLEU* 值见表 5。

表 5 基线模型在结合不同数据增强方法获取到的数据集上的 *BLEU* 值

Table 5 *BLEU* values of baseline NMT model on the datasets augmented by mixed augmentation methods

实验组	en→de	en→fr
1 w Parallel	9.34	9.65
1 w Parallel+1 w BT	14.55	14.86
1 w Parallel+1 w Copy	10.77	11.51
1 w Parallel+1 w BT+2 w WoW	11.14	11.53
1 w Parallel+1 w BT+1 w LDDA-para+1 w LDDA-mono	12.51	13.04
1 w Parallel+1 w Copy+2 w WoW	12.09	13.03
1 w Parallel+1 w Copy+1 w LDDA-para+1 w LDDA-mono	13.49	14.96

由表 5 可见,1 w Parallel+1 w Copy+1 w LDDA-para+1 w LDDA-mono 的数据组合在 en→de 和 en→fr 这 2 个翻译方向上基线神经机器翻译模型都取得了最优的 *BLEU* 指标,与 1 w Parallel+1 w Copy+2 w WoW 的数据组合相比,在 2 个翻译方向的 *BLEU* 指标上分别提升了 1.59 和 1.45。引入 Copy 方法增强的数据能够进一步提升 LDDA 方法的增强效果,因为 Copy 方法将目标语言中的部分命名实体进行了复制,这部分数据在 LDDA 方法抽取出的语料词典中是完全缺失的,从这个维度上对数据进行了增强,即与 LDDA 方法存在互补性。

4 结束语

本文提出了一种提升低资源神经机器翻译性能的数据增强方法 LDDA,并进行了广泛的实验,以验证在不同语言对的低资源场景下,该数据增强方法的有效性。结果表明,与之前的数据增强方法相比,基线 NMT 模型使用经过 LDDA 方法进行数据增强后的数据集进行训练,其 *BLEU* 值得到了显著提高。另外,本文还通过实验证明了 LDDA 可与 Copy 及回译方法相结合,进一步提升基线 NMT 模型的翻译效果。在未来的工作中可尝试在跨领域机器翻译场景下,验证 LDDA 数据增强方法的可用性。

参考文献

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

[2] CURREY A, MICELI-BARONE A V, HEAFIELD K. Copied monolingual data improves low - resource neural machine translation [C]//Proceedings of the Second Conference on Machine Translation. Denmark: dblp, 2017: 148-156.

[3] NAG S, KALE M, LAKSHMINARASIMHAN V, et al. Incorporating bilingual dictionaries for low resource semi - supervised neural machine translation [J]. arXiv preprint arXiv: 2004.02071, 2020.

[4] PENG Wei, HUANG Chongxuan, LI Tianhao, et al. Dictionary-based data augmentation for cross - domain neural machine translation[J]. arXiv preprint arXiv:2004.02577, 2020.

[5] ZHANG Jiajun, ZONG Chengqing. Bridging neural machine translation and bilingual dictionaries [J]. arXiv preprint arXiv: 1610.07272, 2016.

[6] ARTHUR P, NEUBIG G, NAKAMURA S. Incorporating discrete translation lexicons into neural machine translation [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA : dblp, 2016: 1557-1567.

[7] FADAEE M, BISAZZA A, MONZ C. Data augmentation for low-

resource neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: ACL, 2017: 567-573.

[8] WALDENDORF J, BIRCH A, HADOW B, et al. Improving translation of out of vocabulary words using bilingual lexicon induction in low-resource machine translation[C]//Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). ACL, 2022: 144-156.

[9] BURLOT F, YVON F. Using monolingual data in neural machine translation: A systematic study[J]. arXiv preprint arXiv: 1903.11437v1, 2019.

[10] ARTETXE M, LLBAKA G, CASAS N, et al. Do all roads lead to Rome? Understanding the role of initialization in iterative back-translation[J]. Knowledge-Based Systems, 2020, 206: 106401.

[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. Australia: ACL, 2018: 18-24.

[12] ZHENG Zaixiang, ZHOU Hao, HUANG Shujian, et al. Mirror-generative neural machine translation [C]//International Conference on Learning Representations. dblp, 2020: 1-16.

[13] EDUNOV S, OTT M, RANZATO M A, et al. On the evaluation of machine translation systems trained with back-translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 2836-2846.

[14] DOU Z Y, ANASTASOPULOS A, NEUBIG G. Dynamic data selection and weighting for iterative back-translation [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020: 5894-5904.

[15] FADAEE M, MONZ C. Back-translation sampling by targeting difficult words in neural machine translation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL, 2018: 436-446.

[16] YANG Yinfei, ABREGO G H, YUAN S, et al. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax[J]. arXiv preprint arXiv: 1902.08564, 2019.

[17] SCHWENK H. Filtering and mining parallel data in a joint multilingual space[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Australia: ACL, 2018: 228-234.

[18] ARTETXE M, SCHWENK H. Margin-based parallel corpus mining with multilingual sentence embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 3197-3203.

[19] GUPTA K K, SEN S, HAQUE R, et al. Augmenting training data with syntactic phrasal-segments in low-resource neural machine translation[J]. Machine Translation, 2021, 35(4): 661-685.

[20] DYER C, CHAHUNEAU V, SMITH N A. A simple, fast, and effective reparameterization of IBM model 2[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: ACL, 2013: 644-648.

[21] XIAO H. bert-as-service[EB/OL]. [2023-03-01]. https://github.com/hanxiao/bert-as-service.

[22] Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of NAACL-HLT. Minneapolis, Minnesota : ACL, 2019: 4171-4186.

[23] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with GPUS[J]. IEEE Transactions on Big Data, 2019, 7(3): 535-547.