基于单语优先级采样自训练神经机器翻译的研究

张笑燕，逄磊，杜晓峰，陆天波，夏亚梅
（北京邮电大学计算机学院（国家示范性软件学院），北京 100876）

摘　要：为了提高神经机器翻译（NMT）性能，改善不确定性过高的单语在自训练过程中对NMT模型的损害，提出了一种基于优先级采样的自训练神经机器翻译模型。首先，通过引入语法依存分析构建语法依存树并计算单语单词重要程度。然后，构建单语词典并基于单语单词的重要程度和不确定性定义优先级。最后，计算单语优先级并基于优先级进行采样，进而合成平行数据集，作为学生NMT的训练输入。在大规模WMT英德部分数据集上的实验结果表明，所提模型能有效提升NMT的翻译效果，并改善不确定性过高对模型的损害。

关键词：机器翻译；数据增强；自训练；不确定性；语法依存
中图分类号：TP391
文献标志码：A
DOI：10.11959/j.issn.1000−436x.2024066

Research on self-training neural machine translation based on monolingual priority sampling

ZHANG Xiaoyan, PANG Lei, DU Xiaofeng, LU Tianbo, XIA Yamei
School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: To enhance the performance of neural machine translation (NMT) and ameliorate the detrimental impact of high uncertainty in monolingual data during the self-training process, a self-training NMT model based on priority sampling was proposed. Initially, syntactic dependency trees were constructed and the importance of monolingual tokenization was assessed using grammar dependency analysis. Subsequently, a monolingual lexicon was built, and priority was defined based on the importance of monolingual tokenization and uncertainty. Finally, monolingual priorities were computed, and sampling was carried out based on these priorities, consequently generating a synthetic parallel dataset for training the student NMT model. Experimental results on a large-scale subset of the WMT English to German dataset demonstrate that the proposed model effectively enhances NMT translation performance and mitigates the impact of high uncertainty on the model.

Keywords: machine translation, data augmentation, self-training, uncertainty, syntactic dependency

0　引言

利用大规模未标记数据已成为一种提高自然语言处理（NLP, natural language processing）模型效果的有效方法[1]。神经机器翻译（NMT, neural machine translation）是通过深度学习神经网络模型实现的机器翻译方法。平行数据是包含相同内容但使用不同语言编写的文本集合，单语数据是指只包含一种语言的文本数据。NMT作为NLP领域的一个重要分支，需要大量平行数据来训练性能优秀的模型，然而平行数据的匮乏限制了NMT性能的提升。

收稿日期：2024–01–18；修回日期：2024–03–11
通信作者：杜晓峰，dxf@bupt.edu.cn
基金项目：国家自然科学基金资助项目（No.62162060）
Foundation Item：The National Natural Science Foundation of China (No.62162060)

相比平行数据，单语数据的数据量更庞大且容易获得。目前已经有许多方法利用单语数据增强神经机器翻译的性能，比如数据增强[2-4]、半监督训练[5]、预训练[6-7]等。其中，数据增强[2,8]应用最广泛，因为其使用简单且有效。通过数据增强提升模型效果已经成为开发大规模NMT系统的事实标准[9-10]。

作为数据增强的最常用方法之一，自训练[11]能够显著提高神经机器翻译的性能，其通过使用合成的平行数据来增强模型的训练能力。具体而言，自训练包括3个步骤：1) 从大规模单语数据中抽取一个子集；2) 使用一个教师NMT模型将子集数据翻译成目标语言，构建合成的平行数据；3) 将合成的平行数据与真实的平行数据结合，用于训练一个学生NMT模型。自训练有助于降低生成的目标句子的复杂性[12-14]以及提高NMT在低资源语言上的表现[15-16]，那些具有确定性翻译的单语数据不会对自训练的教师NMT模型提供额外的收益[17]。在计算机视觉领域，相关工作也揭示出，在无标签数据中，带有确定性预测的简单模式不会提供额外的收益[18]。最近的研究表明，在后2个步骤中，合成数据的处理[8, 19]和训练策略的优化[20-21]可以显著提升自训练的性能。然而，在第一步中如何高效地从大规模单语数据中抽取子集尚未得到充分研究。

Jiao等[22]提出了一种基于不确定性的自训练神经机器翻译模型，通过真实的平行数据集构建单语词典，计算单语的不确定性，并基于不确定性对单语数据集进行采样，进而合成平行数据集。该方法在一定程度上提高了大规模单语数据的利用率，并提升了NMT模型的性能，但是对于不确定性过高的单语句子，会合成相对较差的翻译结果，这些句子不会提供额外的信息，甚至会阻碍NMT模型的训练。

Duan等[23]在EDA（easy data augmentation）[19]的基础上引入语法依存分析，提出了一种语法感知的数据增强方法。该方法通过构建语法树，选择重要程度较低的单词进行EDA数据增强，补充和提升了EDA的效果，然而该方法仅利用现有平行数据，无法利用大规模的单语数据集，限制了NMT模型性能的进一步提高。

为了解决上述问题，本文提出了一种基于优先级采样的自训练神经机器翻译模型，通过不确定性和语法依存分析定义单语优先级，并基于优先级对大规模单语数据集进行采样。

本文主要的研究工作如下。

1) 在不确定性的基础上，通过引入语法依存分析，定义了单语采样优先级。优先级的定义有效解决了不确定性过高的单语在自训练过程中对NMT模型的损害，同时提高了对大规模单语数据集的利用率。

2) 提出了一种基于优先级的采样策略。通过优先级的定义，进一步探究不同采样策略对NMT的影响，以及基于优先级采样在不同模型上的泛化能力，并设计了基于优先级采样的自训练神经机器翻译模型框架。

3) 在大规模WMT（workshop on machine translation）英德部分数据集上的实验结果表明，基于优先级采样的方式显著提升了NMT模型的翻译效果。NMT模型更多地受益于具有较高优先级的单语句子，同时基于优先级的采样能改善不确定性过高时单语对NMT模型的损害。

1 相关研究

设S代表源语言，T代表目标语言，$\mathcal{X}$代表源语言集合，$\mathcal{Y}$代表目标语言集合，X代表源语言句子，Y代表目标语言句子，其中，$X \in \mathcal{X}$，$Y \in \mathcal{Y}$。另设

$$B = \{(X^i, Y^i)\}_{i=1}^{N} \tag{1}$$

代表真实的平行数据集，其中，N代表平行语句对的数量。设

$$M_x = \{X^j\}_{j=1}^{N_X} \tag{2}$$

代表源语言单语数据集，其中，$X^j \in \mathcal{X}$，N_X代表单语数据集的大小。目标是获得一个翻译模型f: $S \rightarrow T$，能将源语言S翻译成目标语言T。

1.1 不确定性

从源语言到目标语言的翻译可能产生多个不同结果，这就是翻译过程中的不确定性。根据Zhou等[13]的研究，平行数据的复杂程度可以通过累加所有源语言的翻译不确定性来衡量。具体来说，对于源语言X及其翻译候选项目标语言Y，可以计算其条件熵为

$$H(T|S=X) = -\sum_{Y \in \mathcal{Y}} p(Y|X) \log(p(Y|X)) \approx \sum_{t=1}^{T_X} H(y|x = x_t) \tag{3}$$

其中，T_X代表源语言X的长度，x和y分别表示源

语言和目标语言中的一个单词。通常，较高的 $H(T|S=X)$ 表示源语言 X 的复杂程度也较高，同时 X 会有更多的翻译候选项 Y。

式(3)估计了平行数据中源语言 X 与所有可能的翻译候选项 Y 之间的翻译不确定性。然而，式(3)不能直接应用于单语数据中的句子，因为单语数据缺乏对应的翻译候选项。解决这个问题的一种潜在方法是利用一个预先训练好的模型生成多个翻译候选项，这种方法可能会因为生成多样性问题[24-25]而造成偏差估计。更重要的是，对于大规模的单语数据集，生成多个翻译候选项是非常耗时的。

为了解决这个问题，Jiao 等[22]提出了一种基于双语词典衡量单语句子的不确定性方案。具体来说，通过真实的平行数据估计每个源语言单词条件下的目标单词的分布概率，然后使用这个分布来衡量单语句子的翻译不确定性。对于给定的源语言单语句子 $X^j \in M_X$，其不确定性 U 可以表示为

$$U(X^j|V_b) = \frac{1}{T_X}\sum_{t=1}^{T_X} H(y|V_b, x = x_t) \tag{4}$$

在式(4)中，单词级的熵 $H(y|V_b,\ x = x_t)$ 是利用双语词典 V_b 捕捉了每个源单词的翻译方式。双语词典 V_b 记录了每个源单词的所有可能目标单词，以及对应的翻译概率。因此可以通过在真实的平行数据 B 上使用外部对齐工具包来构建词对齐，进而获取 V_b。例如，对于给定的源语言单词 x，有 3 个目标语言单词翻译 y_1、y_2 和 y_3 以及到目标单词的翻译概率 $p(y_1|x)$、$p(y_2|x)$ 和 $p(y_3|x)$，单词级的熵为

$$H(y|V_b, x_i) = -\sum_{y_j \in V_b} p(y_j|x_i)log(p(y_j|x_i)) \tag{5}$$

1.2　语法依存分析

作为自然语言处理的基本任务，语法依存分析旨在预测句子中单词之间的语言依存关系的存在和类型[26-27]。根据在分析树中的搜索策略，依存句法分析器可以大致分为基于图的和基于转移的[28]。随着神经网络在依存分析中应用的发展，语法依存取得了更好的分析性能[26, 29]。Zhang 等[30]提出了一种神经概率依存分析模型，探索了最大似然训练标准下最高三阶基于图的分析。Li 等[31]提出了一个完全基于字符级别的神经依存分析器，并附带了一个针对中文的字符级依存树库。研究表明，依存分析比非神经分析更有效。Wu 等[32]提出了一个用于从原始文本进行多语言通用依存分析的系统。

图 1 展示了“The brown fox jumped over the lazy dog.”的语法依存分析实例。其中，DT、JJ、NN、VBD、IN 等表示词性，det、amod、nsubj、punct 等表示连接关系。这个实例只有一个根节点，而且句子中的每个词都有且只有一个父节点。每个单词和它的父节点之间的标签反映了它们之间的关系。

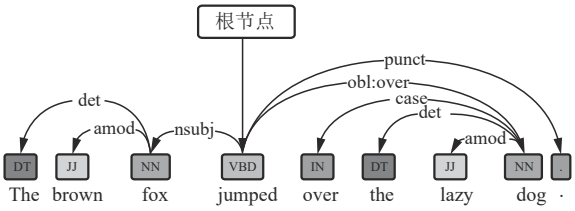

图 1　语法依存分析实例

依存分析树可被视为 NMT 的一个预训练信息，已经被整合到 NMT 中以实现更好的翻译。Eriguchi 等[33]提出了一种树序列模型，其使用基于树的编码器将句子的短语结构编码为向量。Aharoni 等[34]设计了一个序列到树的模型，它将源语言句子翻译成线性化的从属树。具体而言，对于已经存在的双语平行数据，通过语法依存分析筛选出句子中特定的单词，然后进行删除、替换等操作，达到数据增强的目的。该方法使用的前提是存在双语平行数据，因而无法将其应用到单语数据。

2　MPS 自训练模型

2.1　模型概览

当源语言单语不确定性过高时，自训练过程中教师 NMT 会合成错误的目标语言，在这些句子上进行训练会迫使模型过度拟合这些不正确的合成数据，从而导致确认偏见问题[35]。这些结果与先前的研究结果[18, 36]一致，即在某些示例上的学习带来的收益很小，且在极度不确定的示例上进行训练甚至会损害模型。

本文提出了基于 MPS（monolingual priority sampling）的自训练神经机器翻译模型。将语法依存分析引入不确定性，定义单语源语言的优先级，并基于优先级采样进行自训练。具体而言，基于语法依存分析得到源语言单词的重要程度，对单语不确定性进行修正，进而改善不确定性过高对 NMT 模型的损害，提高 NMT 翻译效果。

2.2　优先级

对于给定长度为 n 的源语言句子 X，$X=$

$x_1, x_2, x_3, \cdots, x_n$，其中$x_i$代表第$i$个单词，$i \in [1, n]$，定义$x_i$的重要程度$q_i$为

$$q_i = \frac{1}{2^{d_i - 1}} \tag{6}$$

其中，d_i表示x_i在语法树中的深度，语法树可以通过语法依存分析构建。例如，对于图1所示的语法依存分析，可以构建如图2所示的语法树。

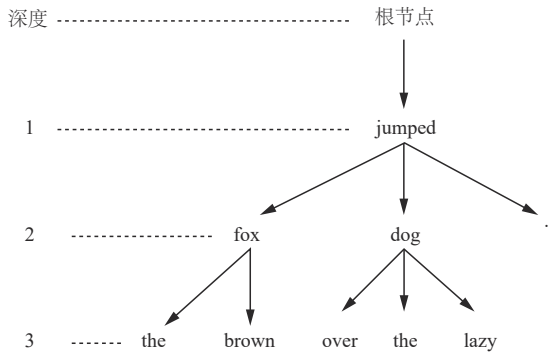

图2　语法树

接下来，对$Q = \{q_i\}$进行softmax归一化处理，得到源语言句子X的重要程度概率分布$P(X)$

$$P(X) = \text{softmax}(Q(X)) = \{p_1, p_2, \cdots, p_n\} \tag{7}$$

例如，对于图1所示句子，其语句重要程度如表1所示。从表1可以看到，当某个单词在语法树中的深度越深，其重要程度p_i越低。

定义单词x_i的优先级$P_r(x_i)$为

$$P_r(x_i) = \frac{H(y|V_b, x = x_i)}{p_i} \tag{8}$$

单语句子X的优先级P_r为

$$P_r = \frac{1}{n}\sum_{i=1}^{n} P_r(x_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{H(y|V_b, x = x_i)}{p_i} \tag{9}$$

其中，$H(y|V_b, x = x_i)$为1.1节中源语言单词的不确定性；p_i为归一化后单词的重要程度，即当单语源语言某个单词x不确定性越高且在整句话中的重要程度越低时，其优先级$P_r(x)$越高，复杂程度越高，生成的合成数据对NMT模型提供的帮助越大，且对模型造成损害的可能性越小。

表1　语句重要程度

单词	d_i	q_i	p_i
the	3	0.25	0.091
brown	3	0.25	0.091
fox	2	0.50	0.117
jumped	1	1.00	0.193
over	3	0.25	0.091
the	3	0.25	0.091
lazy	3	0.25	0.091
dog	2	0.50	0.117
.	2	0.50	0.117

2.3　MPS总体框架

根据上述分析，构建MPS自训练模型整体架构，即选取优先级较高的单语句子作为教师NMT的输入来合成平行数据集，因为这样的句子复杂程度较高，富含更多信息并且对NMT模型的损害较小。MPS自训练模型整体架构如图3所示，其中，S_m表示源语言单语数据集，S_m'表示采样得到的源语言单语数据集，S'和T'表示合成的平行数据集。

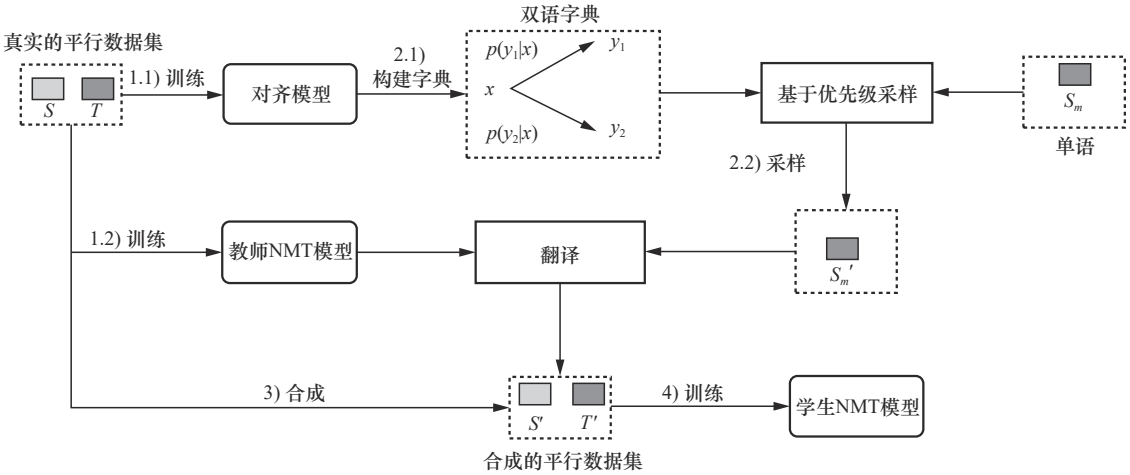

图3　MPS自训练模型整体架构

具体来说主要分为四步：1) 在真实的平行数据集上同时训练一个教师NMT模型和一个对齐模型；2) 从对齐模型中提取双语词典，以及对应的概率分布，并根据优先级按照一定策略进行单语句子的采样；3) 使用教师NMT模型翻译采样得到的单语句子，构建合成的平行数据集；4) 在合成的平行数据集和真实的平行数据集的组合上训练一个学生NMT模型。

3　实验

3.1　实验数据

本文使用了公开的IWSLT（international workshop on spoken language translation）英德部分数据集作为平行数据集，截取部分WMT19 News crawl作为英语单语数据集，其中单语数据集规模大小为平行数据集的30倍以上，该单语数据集相对于平行数据集可以被看作大规模数据集。数据集信息如表2所示。

表2　　数据集信息

数据集类型	大小/MB	语句数量
测试集	0.69	6 750
训练集	18.13	160 250
验证集	0.69	7 284
单语	696.83	4 545 482

训练前首先对数据集进行标点符号规范化处理，清理少数长句、空句以及明显不对齐的句子。此外，数据集都经过BPE（byte pair encoding）[37]处理，实验结果采用由SacreBLEU[38]计算得到的BLEU（bilingual evaluation understudy）[39]值作为评估指标，所有代码均基于PyTorch实现，采用fairseq框架。所有实验使用一块RTX3080显卡，半精度浮点运算（FP16）。

3.2　实验设置

本文模型流程主要分为4个部分，具体描述如下。

1) 训练教师NMT模型和对齐模型。首先在清理好的IWLST英德数据集上训练一个教师NMT模型，本文采用了目前最主流的基于注意力机制的Transformer模型，设置编码器解码器层数为6，前馈层维度为2 048，隐藏层维度为512，注意力为8头。使用fast-align工具训练对齐模型，并建立源语言和目标语言之间单词级别的对齐，这些对齐用于在真实平行数据集中构建双语词典V_b。

2) 计算优先级并采样。首先使用CoreNLP工具对单语数据集进行语法依存分析，然后构建语法树并计算得到单语每个句子中每个单词的重要程度概率分布，根据2.2节可以计算单语数据集中句子的优先级并进行排序，最后对排名较高的R%的单语句子进行采样。

3) 合成平行数据集。将采样得到的源语言单语数据集输入训练好的教师NMT模型中，再将产生的目标语言翻译结果与采样得到的源语言构成合成的平行数据集，最后与真实平行数据集进行合并，得到增强后的平行数据集。

4) 训练学生NMT。在增强后的平行数据集上进行训练，本文分别采用了Transformer、LSTM（long short-term memory）作为学生NMT模型，分别进行验证。其中，LSTM模型采用带注意力机制的LSTM，编码器与解码器隐藏层维度均为512。

3.3　对比及分析

为了验证优先级的有效性，首先分别依据优先级高低和不确定性高低对单语数据集进行排序，然后分别切分成5份，作为教师NMT的输入，进而合成10份增强后的平行数据集，依次作为学生NMT的输入，其中学生NMT采用基于注意力机制的LSTM模型。2种方法在不同数据集上的表现如图4所示。

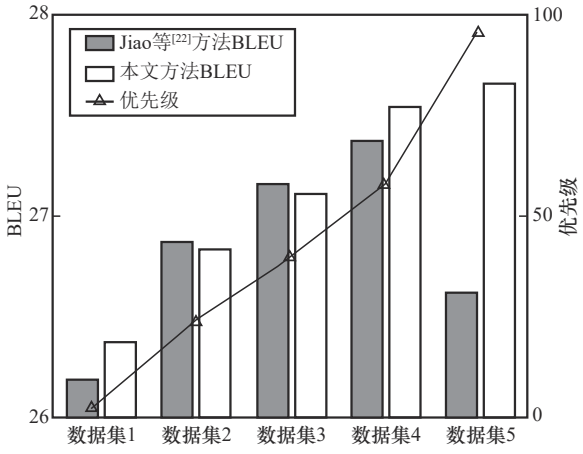

图4　2种方法在不同数据集上的表现

从图4可以看到，随着优先级的提升，模型的翻译效果也在提升，即优先级与模型翻译效果呈线性相关，而在不确定性过高的情况下，模型受到过多错误翻译的损害，BLEU值出现了下降，这与

Jiao等[22]方法结果一致。同时在不确定性和优先级都比较低的情况下（数据集2、3），本文方法较Jiao等[22]方法BLEU值较低，但当不确定性和优先级均较高的情况下（数据集4、5），本文方法效果更好。上述结果说明通过引入语法依存分析判别词的重要程度，本文方法能有效解决不确定性过高对模型的损害，同时在一定情况下提升模型的翻译效果。

为了验证不同采样策略、不同采样率对模型翻译效果的影响，采样策略分别设置为随机采样、优先级采样，采样率R分别为60、70、80、90和100，实验结果如图5所示。

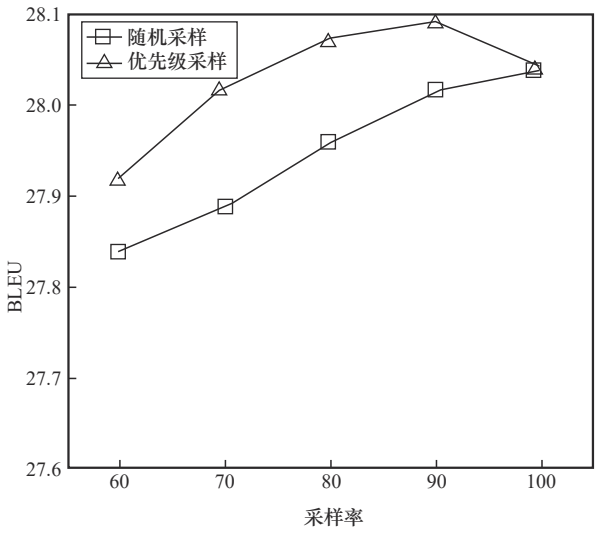

图5　不同采样策略、不同采样率对模型翻译效果的影响

从图5可以看到，当采样率从60增加到90，优先级采样的BLEU值逐步提升，R为90时优先级采样的模型效果最好，但R为100时该模型的BLEU值出现了一定程度的下降。上述结果说明过低的采样率会因为数据量不足对NMT模型的训练造成损失，而过高的采样率则会采集到单语数据集中一些优先级很低的句子，这些句子不会对NMT模型提供额外的信息，甚至会对模型的训练造成损害，影响翻译效果。随机采样的BLEU值随着采样率的提升呈线性增长，且在相同采样率条件下，随机采样的BLEU值明显低于优先级采样的BLEU值，证明高优先级的句子复杂程度也更高，蕴含的信息更多，对模型的训练帮助更大，根据优先级采样对神经机器翻译模型的自训练帮助更大。

为了验证MPS在不同NMT模型上的泛化能力，分别采用Transformer、带注意力的LSTM作为学生NMT模型进行实验，采样率R设置为90。2种模型分别采用Duan等[23]的语法感知数据增强、Jiao等[22]的基于不确定性采样以及本文方法MPS自训练模型，最终结果如表3所示。其中，Real-Text表示原始的真实平行数据集，Dropout[40]表示采取丢弃策略的EDA数据增强，Synthetic（P）和Synthetic（U）分别表示通过优先级采样和不确定性采样得到合成平行数据集再与真实平行数据集合并后的增强数据。

表3　不同模型结果

模型	数据集	BLEU
Transformer（基线）	RealText	24.75
Duan[23]等方法训练的Transformer	RealText + Dropout	27.18
Jiao[22]等方法训练的Transformer	RealText + Synthetic(U)	29.72
MPS（本文方法）训练的Transformer	RealText + Synthetic(P)	**29.94**
LSTM（基线）	RealText	23.17
Duan[23]等方法训练的LSTM	RealText + Dropout	25.62
Jiao[22]等方法训练的LSTM	RealText + Synthetic(U)	28.08
MPS（本文方法）训练的LSTM	RealText + Synthetic(P)	**28.09**

从表3可以看到，在英德翻译任务上，Transformer较基于注意力的LSTM提升了1.58的BLEU值，采用Duan等[23]方法训练的Transformer和LSTM分别较基线提升了2.43和2.45的BLEU值，采用Jiao等[22]方法训练的Transformer和LSTM分别较基线提升了4.97和4.91的BLEU值，而采用MPS分别较基线提升了5.19和4.92的BLEU值。本文方法在Transformer上表现更佳，对模型的提升更大，而在LSTM上带来的提升与Jiao等[22]方法几乎没有差别，这可能是因为LSTM结构导致其更不容易观察到词向量之间的关系。

在2种模型上，Jiao等[22]方法以及本文方法的BLEU值均明显高于Duan等[23]方法，这是因为Duan等[23]方法仅对已经存在的双语平行数据进行扰动，无法利用单语数据。通过有效利用大规模单语数据，其蕴含的信息对模型的训练能提供巨大的帮助。采用本文方法后，2种模型BLEU值较基线

均得到了显著提升，证明本文方法在不同NMT模型上均有效果，具有一定程度上的泛化能力。根据上述结果，本文提出的基于优先级采样的自训练神经机器翻译方法对现有模型可以起到一定的改善作用，具有一定的实用价值。

4　结束语

本文提出了MPS自训练神经机器翻译模型，通过引入语法依存分析、定义优先级，改善了不确定性过高的单语在自训练过程中对NMT模型的损害。通过对比实验证明了本文方法的有效性，并进一步分析了不同采样率、不同采样策略对模型的影响以及不同模型应用本文方法带来的改善。实验表明，在IWLST英德数据集上，本文方法能进一步提升Transformer和LSTM模型的翻译效果，且具有一定的泛化能力。在未来工作中，将进一步优化优先级的定义以及度量，改善低优先级下模型效果较差的问题，并深入探究MPS在不同模型上的表现差异。

参考文献：

[1] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810.04805, 2018.

[2] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[J]. arXiv Preprint, arXiv: 1511.06709, 2015.

[3] PHAM N L, NGUYEN V V, PHAM T V. A data augmentation method for English-Vietnamese neural machine translation[J]. IEEE Access, 2023, 11: 28034-28044.

[4] LAMAR A, KAYA Z. Measuring the impact of data augmentation methods for extremely low-resource NMT[C]//Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023). Stroudsburg: Association for Computational Linguistics, 2023: 101-109.

[5] CAI D, WANG Y, LI H Y, et al. Neural machine translation with monolingual translation memory[J]. arXiv Preprint, arXiv: 2105.11269, 2021.

[6] LIU Y H, GU J T, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. arXiv Preprint, arXiv: 2001.08210, 2020.

[7] VAKHARIA P, VIGNESH S S, BASMATKAR P. Low-resource formality controlled NMT using pre-trained LM[C]//Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Stroudsburg: Association for Computational Linguistics, 2023: 321-329.

[8] EDUNOV S, OTT M, AULI M, et al. Understanding back-translation at scale[J]. arXiv Preprint, arXiv: 1808.09381, 2018.

[9] HASSAN H, AUE A, CHEN C, et al. Achieving human parity on automatic Chinese to English news translation[J]. arXiv Preprint, arXiv: 1803.05567, 2018.

[10] NG N, YEE K, BAEVSKI A, et al. Facebook FAIR's WMT19 news translation task submission[J]. arXiv Preprint, arXiv: 1907.06616, 2019.

[11] ZHANG J J, ZONG C Q. Exploiting source-side monolingual data in neural machine translation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 1535-1545.

[12] KIM Y, RUSH A M. Sequence-level knowledge distillation[J]. arXiv Preprint, arXiv: 1606.07947, 2016.

[13] ZHOU C T, NEUBIG G, GU J T. Understanding knowledge distillation in non-autoregressive machine translation[J]. arXiv Preprint, arXiv: 1911.02727, 2019.

[14] JIAO W X, WANG X, HE S L, et al. Data rejuvenation: exploiting inactive training examples for neural machine translation[J]. arXiv Preprint, arXiv: 2010.02552, 2020.

[15] TONJA A L, KOLESNIKOVA O, GELBUKH A, et al. Low-resource neural machine translation improvement using source-side monolingual data[J]. Applied Sciences, 2023, 13(2): 1201.

[16] XU H Y, WANG X, XING X L, et al. Monolingual denoising with large language models for low-resource machine translation[C]//Proceedings of the International Conference on Natural Language Processing and Chinese Computing. Berlin: Springer, 2023: 413-425.

[17] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 761-769.

[18] MUKHERJEE S, AWADALLAH A H. Uncertainty-aware self-training for text classification with few labels[J]. arXiv Preprint, arXiv: 2006.15315, 2020.

[19] CASWELL I, CHELBA C, GRANGIER D. Tagged back-translation[J]. arXiv Preprint, arXiv: 1906.06442, 2019.

[20] WU L J, WANG Y R, XIA Y C, et al. Exploiting monolingual data at scale for neural machine translation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 4207-4216.

[21] WANG S, LIU Y, WANG C, et al. Improving back-translation with uncertainty-based confidence estimation[J]. arXiv Preprint, arXiv: 1909.00157, 2019.

[22] JIAO W X, WANG X, TU Z P, et al. Self-training sampling with monolingual data uncertainty for neural machine translation[J]. arXiv Preprint, arXiv: 2106.00941, 2021.

[23] DUAN S F, ZHAO H, ZHANG D D. Syntax-aware data augmentation for neural machine translation[J]. ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2988-2999.

[24] LI J W, MONROE W, JURAFSKY D. A simple, fast diverse decoding algorithm for neural generation[J]. arXiv Preprint, arXiv: 1611.08562, 2016.

[25] SHU R, NAKAYAMA H, CHO K. Generating diverse translations

with sentence codes[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 1823-1827.

[26] LI Z C, HE S X, ZHANG Z S, et al. Joint learning of POS and dependencies for multilingual universal dependency parsing[C]//Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Stroudsburg: Association for Computational Linguistics, 2018: 65-73.

[27] HE S X, LI Z C, ZHAO H, et al. Syntax for semantic role labeling, to be, or not to be[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 2061-2071.

[28] LI Z C, CAI J X, HE S X, et al. Seq2seq dependency parsing[C]//Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2018: 3203-3214.

[29] XIE Z A, WANG S, LI J W, et al. Data noising as smoothing in neural network language models[J]. arXiv Preprint, arXiv: 1703.02573, 2017.

[30] ZHANG Z S, ZHAO H, QIN L H. Probabilistic graph-based dependency parsing with convolutional neural network[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2016: 1382-1392.

[31] LI Z C, CAI J X, ZHAO H. Effective representation for easy-first dependency parsing[C]//Pacific Rim International Conference on Artificial Intelligence. Berlin: Springer, 2019: 351-363.

[32] WU Y T, ZHAO H, TONG J J. Multilingual universal dependency parsing from raw text with low-resource language enhancement[C]//Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Stroudsburg: Association for Computational Linguistics, 2018: 74-80.

[33] ERIGUCHI A, HASHIMOTO K, TSURUOKA Y. Tree-to-sequence attentional neural machine translation[J]. arXiv Preprint, arXiv: 1603.06075, 2016.

[34] AHARONI R, GOLDBERG Y. Towards string-to-tree neural machine translation[J]. arXiv Preprint, arXiv: 1704.04743, 2017.

[35] ARAZO E, ORTEGO D, ALBERT P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning[C]//Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-8.

[36] CHANG H S, LEARNED-MILLER E, MCCALLUM A. Active bias: training more accurate neural networks by emphasizing high variance samples[J]. arXiv Preprint, arXiv: 1704.07433, 2017.

[37] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[J]. arXiv Preprint, arXiv: 1508.07909, 2015.

[38] POST M. A call for clarity in reporting BLEU scores[J]. arXiv Preprint, arXiv: 1804.08771, 2018.

[39] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM Press, 2002: 311-318.

[40] WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv Preprint, arXiv: 1901.11196, 2019.

[作者简介]

张笑燕（1973– ），女，山东烟台人，博士，北京邮电大学教授，主要研究方向为软件工程理论、移动互联网软件与大数据分析。

逄磊（1999– ），男，山东青岛人，北京邮电大学硕士生，主要研究方向为自然语言处理、机器翻译等。

杜晓峰（1973– ），男，陕西韩城人，北京邮电大学讲师，主要研究方向为云计算与大数据分析。

陆天波（1977– ），男，贵州毕节人，博士，北京邮电大学教授，主要研究方向为网络与信息安全、安全软件工程和P2P计算。

夏亚梅（1976– ），女，甘肃天水人，博士，北京邮电大学副教授，主要研究方向为网络安全与数据挖掘。