计算机科学
COMPUTER SCIENCE

基于长尾词分布的藏汉机器翻译数据增强方法

格桑加措, 尼玛扎西, 群诺, 嘎玛扎西, 道吉扎西, 罗桑益西, 拉毛吉, 钱木吉

引用本文

格桑加措, 尼玛扎西, 群诺, 嘎玛扎西, 道吉扎西, 罗桑益西, 拉毛吉, 钱木吉. 基于长尾词分布的藏汉机器翻译数据增强方法[J]. 计算机科学, 2026, 53(1): 224-230.

KALZANG Gyatso, NYIMA Tashi, QUN Nuo, GAMA Tashi, DORJE Tashi, LOBSANG Yeshi, LHAMO Kyi, ZOM Kyi. Data Augmentation Methods for Tibetan-Chinese Machine Translation Based on Long-tail Words [J]. Computer Science, 2026, 53(1): 224-230.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200147

基于长尾词分布的藏汉机器翻译数据增强方法

格桑加措　尼玛扎西　群　诺　嘎玛扎西　道吉扎西　罗桑益西　拉毛吉　钱木吉
西藏大学信息科学技术学院　拉萨 850000
西藏大学藏文信息技术教育部工程研究中心　拉萨 850000
(gyatso736@outlook.com)

摘　要　现有藏汉机器翻译语料中存在领域数据分布不平衡的问题,导致训练出来的模型对各个领域数据的翻译能力表现不均衡。反向翻译作为一种常见的数据增强方法,通过提供更多样化的伪数据来提高模型的性能。然而,传统的反向翻译方法难以充分考虑数据的领域分布不平衡问题,导致模型在整体性能提升过程中难以提升资源稀缺领域的翻译性能。对此,通过深入分析语料中的长尾词的分布,有针对性地利用现有藏汉双语语料的长尾词来选取单语数据,通过反向翻译构造伪数据进行数据增强操作。这一策略旨在提升藏汉机器翻译模型整体性能的同时,改善数据匮乏领域的翻译性能。实验结果表明,通过充分考虑领域数据不平衡情况,结合长尾词数据增强,能够有效提升机器翻译模型在稀缺领域的翻译性能,为解决领域数据不平衡问题提供了一种有针对性的策略。

关键词:长尾词;数据增强;藏汉机器翻译;领域数据不平衡
中图分类号　G35;TP391

Data Augmentation Methods for Tibetan-Chinese Machine Translation Based on Long-tail Words

KALZANG Gyatso,NYIMA Tashi,QUN Nuo,GAMA Tashi,DORJE Tashi,LOBSANG Yeshi,LHAMO Kyi and ZOM Kyi
School of Information Science and Technology,Tibet University,Lhasa 850000,China
Tibetan Language Information Technology Ministry of Education Engineering Research Center of Tibet University,Lhasa 850000,China

Abstract　The existing Tibetan-Chinese machine translation corpora exhibit significant domain data imbalance,resulting in inconsistent translation performance of trained models across different domains. Back-translation,as a common data augmentation method,enhances model performance by generating diverse pseudo-parallel data. However,traditional back-translation approaches struggle to fully account for the domain imbalance in the data distribution,leading to limited improvements in translation performance for resource-scarce domains,even as overall performance increases. This paper proposes a strategy that involves an in-depth analysis of the distribution of long-tail words in existing corpora,and targeted selection of monolingual data using these long-tail words from the existing Tibetan-Chinese bilingual corpora. By generating pseudo-data through back-translation,it performs data augmentation. This strategy aims to improve the overall performance of Tibetan-Chinese machine translation models while enhancing translation performance in data-scarce domains. Experiment results demonstrate that by fully considering domain data imba-lance and incorporating long-tail word data augmentation,the translation performance of machine translation models in resource-scarce domains can be effectively improved,providing a targeted approach to address the issue of domain data imbalance.

Keywords　Long-tail words,Data augmentation,Tibetan-Chinese machine translation,Domain data imbalance

随着深度学习技术的快速发展,自然语言处理领域取得了令人瞩目的进步[1-5],这主要得益于大规模数据的积累和算法的持续优化。然而,在现实应用中,数据通常存在类别分布不均的现象[6],即数据的长尾特性。长尾特性表现为某些类别的数据占据整体数据分布的大部分,而其他类别的数据则仅占很小比例(见图1),导致数据分布不均匀。这给在特定任务中对低频类别进行有效学习带来了严峻挑战。在机器翻译领域,双语句对的质量、规模和领域分布是决定翻译模型性能的关键因素[7-10]。现有的藏汉双语句对中存在明显的领域数据分布不平衡。以民俗、宗教和传统医学等特殊领域为例,

到稿日期:2024-12-18　返修日期:2025-03-07
基金项目:新一代人工智能国家科技重大专项(2022ZD0116101);国家自然科学基金重点项目(62436006);西藏自治区自然科学基金重点项目(XZ202401ZR0040);西藏自治区自然基金面上项目(XZ202401ZR0031);国家自然科学基金青年基金(62406257)
This work was supported by the National Key R&D Program of the New Generation of Artificial Intelligence(2022ZD0116101),Key Project of the National Natural Science Foundation of China(62436006),Key Project of Xizang Natural Science Foundation(XZ202401ZR0040),General Project of Xizang Natural Science Foundation(XZ202401ZR0031) and Youth Fund of National Natural Science Foundation of China(62406257).
通信作者:尼玛扎西(nmzx@utibet.edu.com)

这些领域包含大量低频的专业词汇和术语，通常位于长尾分布的末端。由于这些低频词汇在语料库中的稀缺性，翻译模型难以对其进行有效学习和泛化，因此其在相关领域的翻译性能较低。

领域数据不平衡指的是在训练数据中，不同领域或类别样本分布明显不均的情况[11]。具体而言，某些领域的数据量远远超过其他领域，模型在训练过程中更倾向于适应数据丰富的领域，而对数据稀缺的领域表现不佳。这种数据分布问题导致模型在面向特定领域时性能下降，因为模型未能充分接触并学习稀缺领域的知识。在藏汉双语任务中，类似问题尤为突出。例如，在藏医药和佛学领域，专业术语稀缺性限制了模型的学习效果。这些问题不仅涉及长尾词汇的低频分布，还面临缺乏针对特殊领域标注数据的额外挑战。

解决领域数据不平衡问题需要采用有效的数据平衡策略，以确保模型能够全面学习各领域的特性。然而，由于特定领域高质量平行语料的匮乏，机器翻译系统在这些领域的表现往往不理想。这是藏汉低资源领域机器翻译面临的核心挑战之一。现有语料主要集中在新闻和时事等通用领域，藏医药、佛学等领域的数据尤为稀缺。这种不均衡的分布导致模型在训练时偏向对数据丰富领域知识进行学习，使得在测试集中领域分布均衡的条件下模型性能显著下降，如图 2 所示。

此外，当前翻译模型通常采用基于 softmax 的交叉熵损失函数，其未能充分考虑领域数据的不平衡性。交叉熵损失函数将一个类别的每个正样本视为其他类别的负样本，使得数据丰富类别接收到更强的梯度信号，而数据稀缺类别仅获得较弱的梯度更新。这种梯度更新的不平衡进一步加剧了领域数据的不均衡问题，削弱了模型对低频类别的学习能力。

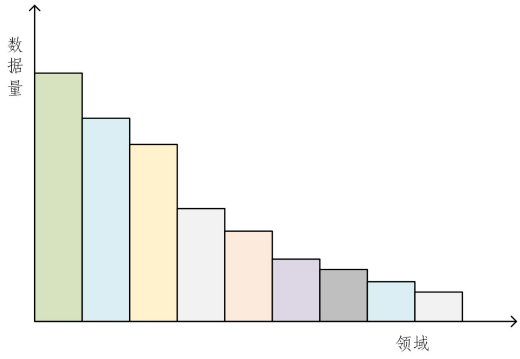


图 1 训练集中的领域数据分布图直观图

Fig. 1 Visual distribution of domain data in training set

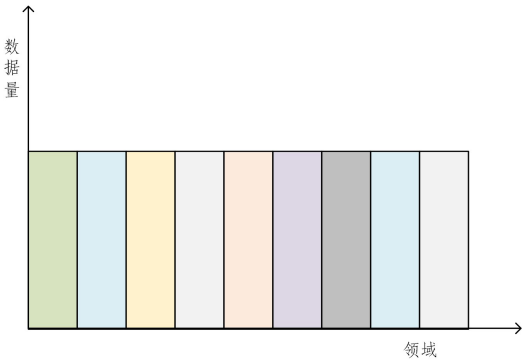


图 2 测试集中的领域数据分布图直观图

Fig. 2 Visual distribution of domain data in test set

针对领域数据不平衡问题，研究者提出了多种改进方法，例如重采样[12]、过采样[13]以及迁移学习[14-15]等，旨在提高模型在稀缺数据领域的性能。数据增强策略[16-19]被广泛认为是提升模型性能的有效途径之一，用于增加训练数据的多样性，从而提高模型的性能和泛化能力，被广泛应用于计算机视觉、自然语言处理等领域。其基本思想是通过对原始训练数据进行一系列变换，生成具有相似但不完全相同特征的新样本。该策略可以有效扩大训练集，使模型能够在更多场景下学习到有用的特征，同时降低过拟合的风险。在计算机视觉中，常见的数据增强方法包括图像的旋转、翻转、缩放、裁剪和平移等。在自然语言处理中，数据增强方法通常包括同义词替换、随机插入、随机删除、随机打乱词序以及改变句子结构等。

在机器翻译领域，数据增强技术有助于提高模型对缺乏双语句对的领域数据的泛化能力，减少对特定训练样本的过度依赖，使模型能够更好地适应未见过的数据，特别是通过生成伪数据来扩充稀缺领域的训练样本[20-22]。然而，现有方法在稀缺领域中的表现仍存在局限性，其未能充分结合领域数据不平衡的特点，导致生成的数据难以显著改善特定领域的翻译效果。

根据 Zipf 定律[23]，尽管长尾词出现频率较低，但在特殊领域中往往承担着重要的语义功能。分析长尾词的分布有助于揭示数据集中的领域覆盖情况，并为模型训练提供更有针对性的指导。基于此，本文提出了一种利用长尾词进行数据增强的方法。通过对 Zipf 曲线中长尾词分布的详细分析，有针对性地利用现有语料中的长尾词结合反向翻译方法，来增加稀缺领域的训练数据。这一策略不仅显著提升了模型的整体翻译性能，还有效改善了模型在数据稀缺领域的泛化能力。研究结果表明，结合长尾词的数据增强方法能够为解决领域数据不平衡问题提供一种可行且高效的策略，同时拓展了数据增强在机器翻译领域的应用前景。

1 Zipf 分布及长尾词

1.1 Zipf 分布

Zipf 定律[22]由美国语言学家 George Zipf 于 1932 年提出，他在研究英文单词出现频率时发现，当将单词按出现频率从高到低排序时，单词的频率 f 与其排名 r 呈现出一种反比关系，如式(1)所示：

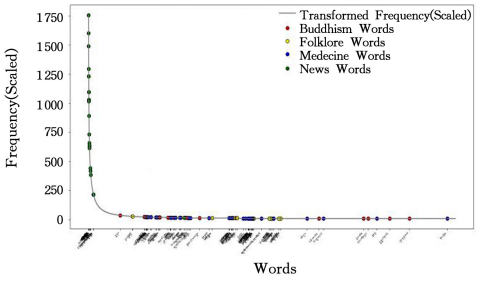
$$f \propto \frac{1}{r^s} \tag{1}$$

其中，s 通常接近于 1。

该现象不仅存在于自然语言的词频分布中，还广泛适用于其他领域的数据分布，展示了 Zipf 分布的普适性与重要性。在语言学研究中，Zipf 定律反映了语言表达的经济性和效率。高频词(如“the”或“and”)虽然种类较少，但在交流中起到了关键的连接作用，而低频词则丰富了语言的表达能力。语言系统的这种高效分布通常被认为是人类语言优化的结果。

为进一步分析藏语在特定领域中的词汇分布规律，随机挑选了佛学、医学、传统文学和新闻 4 个领域的词汇各 30 余条，分别用绿色(新闻)、红色(佛学)、黄色(传统文学)、蓝色

(医学)进行散点可视化展示,如图 3 所示。从可视化结果可以看出,4 个领域的词汇分布总体上符合长尾分布规律:新闻领域的高频词汇主要分布在 Zipf 曲线的头部区域,而佛学、医学和传统文学的词汇更多地集中在尾部。这一现象与我们的预期一致,进一步佐证了现有藏汉语料中不同领域数据的词频分布规律存在显著差异,且在领域特定语料中表现出自然语言的长尾特性。该发现为后续基于长尾词的数据增强方法奠定了理论基础。



注:Buddhism words 指佛学领域词,Folklore words 指传统藏学领域词,Medecine words 指医学领域词,News words 指新闻领域词。

图 3 现有语料中藏语词典的词频分布曲线图

Fig. 3 Word frequency distribution curve of Tibetan dictionary in existing corpus

1.2 长尾词

长尾词[24-25]是指在语料中出现频率较低,但可能承载特定领域或语境中重要语义信息的词汇。这类词语通常位于词频分布的尾部,是 Zipf 分布中的一个重要组成部分。与高频词不同,长尾词虽然单独出现的次数有限,但其集合的总量却极为庞大,在语料中呈现出“少数高频词和大量低频词”的分布模式。在现有的藏汉双语翻译语料中,各领域数据存在明显的长尾词分布现象,如图 4 所示。

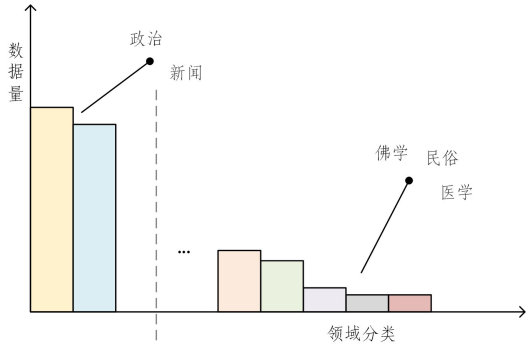


图 4 现有藏汉双语语料中的领域分布直观图

Fig. 4 Domain distribution visualization in the existing Tibetan-Chinese bilingual corpus

在特定主题或领域中,长尾词往往与关键的领域知识相关。例如,在医学领域的语料中,“心肌梗塞”或“冠状动脉”这样的低频词可能是核心术语,而在法律文本中,“合同解除”或“违约责任”等词汇具有重要的法律语义。这些词语在一般语言模型中可能被忽视,但在领域任务中却承担着不可替代的作用。

长尾词在特定主题或领域中往往承担着关键的语义任务,因此,对语料中的长尾词分布进行系统分析,重点探讨其在领域数据不平衡问题[25]中的作用。首先,统计了长尾词的频率分布及其占总语料词汇量的比例,并对不同领域的长尾词特性进行了比较,发现长尾词在特定领域语料中的分布规律符合长尾效应,且这些低频词在领域相关的翻译任务中具有不可忽视的贡献。充分利用这些长尾词进行数据增强,不仅可以提升模型在特定领域的翻译性能,还能在一定程度上提升模型的整体表现。虽然这种方法的效果可能因具体情况而异,但本文研究结果表明,基于长尾词的针对性数据增强策略是解决领域数据不平衡问题的一种有效途径。

2 基于长尾词的数据增强

2.1 长尾词的选择及单语数据的抽取

为了深入探讨领域数据的不平衡问题,本文通过分析整体词典的 Zipf 词语分布,从中选择了位于长尾分布尾部的低频词汇进行研究。具体而言,统计了在长尾分布中出现次数小于 5 次的汉语词汇,并基于这些长尾词从真实语料库中提取了包含这些词汇的汉语句子。这样做的目的是为了捕捉在常规语料中较为稀有,但在特定领域中可能具有重要语义作用的词汇,这些词汇通常承担着领域特有的知识传递和语义表达功能。

本文使用了由 CCMT2023 提供的超过 500 万条单语数据。为了确保数据的质量和有效性,对这些数据进行了去重和去除过长或过短的句子等全面的预处理,得到的 330 多万条质量较高的句子用于后续实验的单语数据支撑,进行反向翻译。

2.2 反向翻译模型

反向翻译[26]是数据增强中最常用的方法之一,其将目标语言的单语数据翻译回源语言,从而构造双语句对。其动机在于充分利用单语数据的易得性,通过反向翻译的方式提升翻译质量,进而提高翻译系统的整体性能。反向翻译在机器翻译领域得到了广泛关注,特别是在现有语料受限、无法充分捕捉特定语境中的语义和语法信息时。通过反向翻译,可以利用已有资源,基于单语语料引入更多语境和语义信息,改进翻译系统在特定领域和语境下的表现。作为一种有效的数据增强手段,反向翻译通过引入更多训练样本,提高了翻译系统的泛化能力,进而提升了翻译结果的准确性。

在本文实验中,首先在原始的 140 多万条藏汉双语平行语料基础上,利用 Transformer(Base)模型训练汉语到藏语的反向翻译模型。该模型为后续的单语数据生成伪双语数据提供了基础。之后,对 CCMT2023 提供的 500 多万条汉语单语数据进行预处理,挑选出 330 多万条较为规范的汉语单语句子。随后利用训练好的汉藏反向翻译模型,将这些汉语单语句子翻译为对应的藏语句子,生成相应的伪双语句对。

2.3 总双语句对构造

在完成单语数据和长尾词的选取及预处理后,进一步构造了总双语句对。为了有效结合长尾词的特性,本文采用

Similarities 工具[1)]，并结合中文词向量模型，对 330 多万条伪双语句对进行筛选。本文的目标是从这些伪句对中找出与长尾词相关性较高的句子，从而增强模型在稀缺领域的翻译能力。具体而言，首先利用词向量模型计算每个句子的向量表示，并基于这些向量计算句子之间的相似性。通过这种方式，可以量化句子与长尾词的相关度。接下来，根据计算得到的相似性分数对句子进行排序，并从高到低依次筛选出最相关的句子。这些句子中，与长尾词的匹配度较高的句子被优先选出，并组成不同数量的子集，具体为 30 万条、50 万条、100 万条和 150 万条句子。

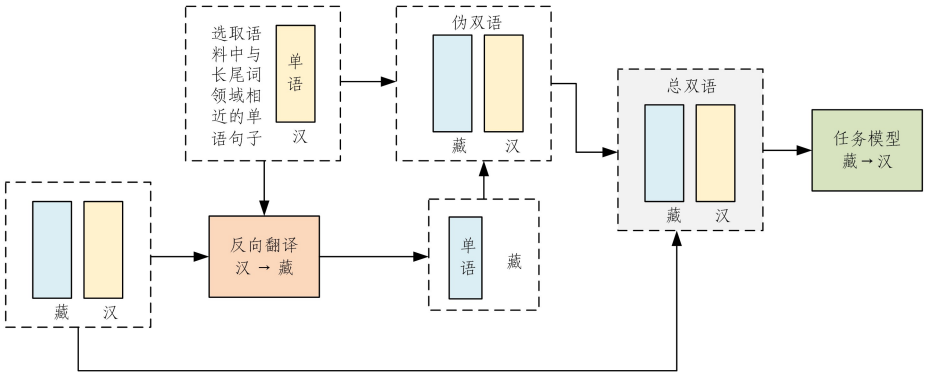


图 5 基于长尾词的数据增强方法的整体流程图

Fig. 5 Overall process flow of data augmentation based on long-tail words

为了确保这些句子具有高质量的翻译信息，使用前述反向翻译方法对选出的句子进行翻译，从而生成新的伪双语句对。通过这种方式，原本单一的单语数据被转换为高质量的双语数据，有效扩充了训练数据集。

最终，将这些通过反向翻译生成的伪双语句对与原有的真实句对结合，形成一个包含大规模双语句对的训练集。这个增强后的训练集不仅包含了大量的高质量句子，还有效地增加了长尾词的覆盖度，为后续的机器翻译模型训练提供了更为丰富的数据支持。通过这一过程，本文重新训练了机器翻译模型，并在新的数据集上进行测试，进一步提升了模型在特定领域的翻译性能。整体流程如图 5 所示。

由于本文收集到的单语句子数量只比真实双语句对多两倍多，加上未能针对特殊语料构造伪双语句对，一定程度上限制了本文方法的性能表现。最终，本文使用一系列测试集对模型进行了评估，并将实验结果与基于随机挑选的单语数据增强方法进行了比较。

3 实验设置及模型选择

3.1 实验环境

本文所有模型均在一台 NVIDIA GeForce RTX 3080 显卡上进行训练。所有实验均基于 Transformer[1] 模型进行训练，模型参数包括编码器和解码器，它们各包含 6 个隐藏层，学习率设定为 0.015，每个模型训练 25 轮。

3.2 实验数据

实验所用训练数据集来源于西藏大学和 CCMT 2023 年的数据，包含超过 140 条藏汉双语句对。其中，西藏大学的数据包含少量医学、佛学和口语等领域的内容，经过预处理后，从中抽取了 1 万条作为验证集，剩下的 130 多万条则作为训练集。本文所使用的单语数据来源于 CCMT2023 汉语单语数据集，在经过一系列预处理后，最终保留了 330 多万条数据。

为了全面评估翻译模型在不同领域的性能，测试集分为 5 个部分：多领域测试集，包含藏医、佛学、口语（包括电影翻译口语和日常口语两部分）的测试集和 CCMT2020 验证集。每个领域各包含 250 条句对，共 1 000 条句对。CCMT2020 验证集作为通用领域的测试集；医学药领域测试集 700 条句对；日常口语领域测试集 500 条句对；电影口语（汉语电影翻译成藏语的翻译口语）测试集 500 条。这些数据集不仅涵盖了日常用语，还包括了专业术语和特定领域的翻译，为机器翻译基础模型的训练和性能评估提供了丰富的语料资源。通过在这 5 种不同的测试集上进行评估，旨在深入了解模型在处理不同领域特性翻译任务时的表现。这种全面的评估有助于了解模型的领域适应性和整体性能，为进一步优化和改进模型提供依据。

3.3 模型及相对位置编码

Transformer[1] 模型由 Vaswani 等在 2017 年提出，其核心是自注意力机制，这种机制允许模型在处理输入序列时关注序列中的不同位置。通过计算注意力权重，模型能够将输入序列中的每个元素关联到输出序列的每个位置，而不仅仅依赖固定的权重。这种机制使得每个输入位置可以动态调整对输出的贡献，从而提高了模型的灵活性和表达能力。

Transformer 模型使用多个注意力头（Multi-head Attention），每个头学习不同的注意力权重，这些头的输出被拼接在一起，并通过线性变换得到最终的输出。然而，由于 Transformer 本身不包含序列的位置信息，因此模型引入了位置编码来帮助学习序列的顺序信息。Transformer 模型由编码器和解码器两部分组成。编码器负责将输入序列编码成一个上下文表示，解码器则使用这个上下文表示生成目标序列。在机器翻译任务中，源语言句子经过编码器处理生成上下文表示，解码器再根据该上下文表示生成目标语言句子。每个子层（如注意力机制、前馈神经网络等）后面都添加了残差连接，以便更好地训练深层网络。此外，层归一化用于稳定训练

1) https://github.com/shibing624/similarities

过程。编码器和解码器中的前馈神经网络在每个位置独立地处理序列元素，增强了模型的非线性建模能力。Transformer-base是目前最通用的模型之一，因此选择其作为基础模型框架，其参数如表1所列。

表1 Transformer(Base)参数

Table 1 Transformer(Base) parameters

Transformer Base	参数
Hiden_size	512
Feed-Forward filter_size	2048
num_heads	8
num_encoder_layers	6
num_decoder_layers	6

考虑到词的位置对实验结果的影响，在基础模型上应用了相对位置编码的改进。相对位置编码[21]是针对原始模型Transformer中使用的绝对位置编码无法捕捉序列之间的相对位置关系而提出来的改进方法。在机器翻译任务中，相对位置编码的改进能提升模型的翻译性能。本文进行了位置编码改进前后的翻译性能对比实验，并在所用测试集上进行了验证，如表2所列。

表2 相对位置编码改进前后的模型性能对比

Table 2 Comparison of model performance before and after the improvement of relative positional encoding

编码	*BLEU* 值			
	通用领域	现代医学	佛学	藏医
绝对位置编码	49.49	13.18	2.24	0
相对位置编码	52.32↑	12.75	3.86↑	0

注：↑表示性能有所提升。

结果表明，除现代医学领域外，其他测试集均在相对位置编码改进后实现了稳步的性能提升。特别是在通用领域测试集上，性能提升最为显著。因此，所有实验均在改进后的相对位置编码框架下进行训练。

3.4 分词系统

对于分词系统，藏文文本使用西藏大学基于Bi-LSTM和CRF结合的藏文分词系统[1)]进行分词，由于针对未登录音节词进行了模型优化，适用于处理大量的藏语文本数据，准确度较高。对于汉文文本，本文使用jieba分词器[2)]进行分词处理。

3.5 评测指标

本文实验选用BLEU值作为评估指标，并以4-gram为基础进行计算。BLEU由Papineni等[27]于2002年提出，是一种常用于机器翻译(Machine Translation，MT)评估的指标，用于衡量机器生成的翻译结果与参考翻译之间的相似程度。BLEU值的计算基于n-gram的匹配和一个惩罚因子，其中惩罚因子会考虑短句子的问题，避免因为翻译的句子较短而得到较高的值。BLEU值的范围为0～1，越接近1表示机器翻译结果越好。

4 实验结果及分析

4.1 实验结果

为了验证本文方法的可靠性，在原始的140多万条双语数据基础上，通过随机抽取方法和本文方法，分别从330万条单语数据中选取50万条、100万条和200万条数据进行反向翻译。随后，将生成的伪双语数据与原始数据融合，用于模型训练。为了评估模型性能，对测试数据进行了交叉验证，实验结果如表3所列。

表3 同数据不同模型在测试数据上的BLEU值

Table 3 BLEU scores of different models on test data with different datasets

(%)

数据	方法	*BLEU* 值				
		医学	日常口语	电影口语	多领域	2020验证集
原始数据	—	12.75	19.35	11.08	11.45	52.32
+30万条	随机挑选	14.91	19.66	11.53	12.34	50.94
(伪数据)	本文方法	15.45↑	19.87↑	12.95↑	12.94↑	51.52↑
+50万条	随机挑选	15.81	19.31	11.25	11.97	51.40
(伪数据)	本文方法	15.88↑	19.79↑	13.06↑	12.68↑	50.94
+100万条	随机挑选	16.97	19.29	13.39	11.84	49.83
(伪数据)	本文方法	16.71	19.45↑	13.69↑	12.67↑	50.33↑
+150万条	随机挑选	17.34	18.88	13.80	11.98	49.21
(伪数据)	本文方法	17.57↑	20.52↑	14.91↑	13.63↑	49.42↑

注：↑代表性能有所提升。

4.2 结果分析

实验结果表明，在相同数据量条件下，本文方法能够在日常口语测试集、电影口语测试集和多领域测试集上显著提升BLEU值，与随机抽取方法相比效果更佳。具体来看，在多领域测试集中，分别添加30万条、50万条、100万条和150万条伪数据时，BLEU值分别提升了0.6个百分点、0.71个百分点、0.83个百分点和1.65个百分点。当添加150万条伪数据时，BLEU值达到最高，为13.63%，较原始数据提升了2.18个百分点。在电影口语测试集中，添加相同数量的伪数据后，BLEU值分别提升了1.42个百分点、1.81个百分点、0.3个百分点和1.11个百分点，150万条伪数据下的BLEU值达到14.91%，比原始数据提升了3.83个百分点。日常口语测试集的BLEU值则分别提升了0.21个百分点、0.48个百分点、0.16个百分点和1.64个百分点，在添加150万条伪数据时，BLEU值达到最高，为20.51%，比原始数据提升了1.17个百分点。总体来看，随着伪数据量的增加，本文方法在各个测试

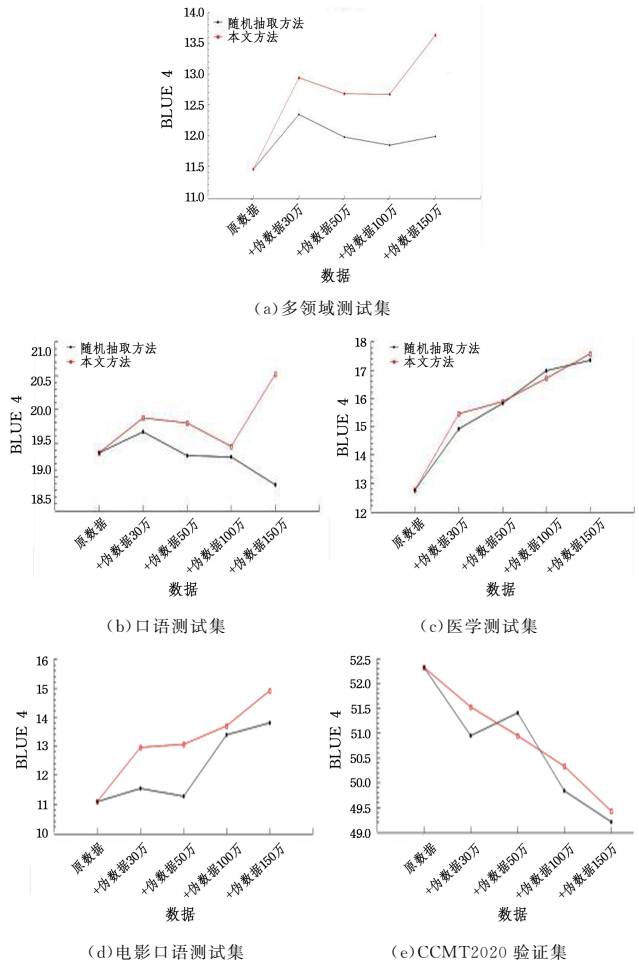
1) https://github.com/gyatso736/-Tibetan-tokenizer-

2) https://github.com/fxsjy/jieba

集上的 BLEU 值都表现出稳步提升的趋势，尤其是在多领域和电影口语测试集上，效果显著。这一结果验证了本文方法在处理不同领域翻译任务中的有效性，尤其是在数据稀缺领域的性能提升方面，提供了切实可行的解决方案。

对于医学领域测试集，当添加 30 万、50 万和 150 万条伪数据时，BLEU 值分别提升了 0.54 个百分点、0.07 个百分点和 0.23 个百分点。然而，在添加 100 万条伪数据时，BLEU 值反而下降了 0.26 个百分点。对于 2020 验证集，虽然在添加伪数据后，BLEU 值相较于原始数据有所下降，但使用本文方法在添加 30 万、100 万和 150 万条伪数据时，BLEU 值均高于随机抽取方法。这表明，尽管在某些情况下添加伪数据可能导致性能下降，但本文方法在整体上仍然优于随机抽取的方法，尤其在应对领域数据不平衡的问题时展现出了较好的效果。

对两种方法进行了性能可视化，从图 6 中可以看出，本文方法在多领域测试集、口语（日常）测试集、医学测试集和电影口语（翻译口语）测试集上的表现普遍优于传统的反向翻译方法。



(a)多领域测试集　(b)口语测试集　(c)医学测试集　(d)电影口语测试集　(e)CCMT2020 验证集

图 6　两种方法在各个测试集上的性能可视化图

Fig. 6　Visualization of performance of two methods on different test sets

在 CCMT2020 验证集上的性能经过反向翻译后有所下降，主要原因是 CCMT2020 测试集的领域分布与原始数据非常接近，原始数据上的 BLEU 值达到了 52.32%。对该领域来说，反向翻译生成的伪双语数据反而引入了更多噪声，导致性能下降。然而，从 CCMT2020 测试集上的结果可视化图（见图 6(e)）可以看出，除了在添加 50 万条伪数据的情况下性能略有下降之外，本文方法在其他伪数据量级上依然表现出优于传统反向翻译方法的性能。

结束语　本文研究了藏汉机器翻译系统在应对领域数据不平衡问题时面临的挑战和改进的策略。尽管现有的藏汉机器翻译系统在新闻、政治等资源较为丰富的领域已能达到令人满意的翻译效果，能够满足日常翻译需求，但在医学、口语、佛学等资源匮乏的特殊领域，由于缺乏相关领域的双语句对且建立语料的成本高昂，因此训练出的模型在这些领域的翻译性能明显不佳。领域数据不平衡问题已成为当前藏汉机器翻译任务中亟待解决的难题。

为解决这一问题，本文提出了一种基于现有语料中 Zipf 分布长尾词特性的策略。通过有针对性地选取包含长尾词的单语数据进行数据增强操作，旨在提升模型在稀缺数据领域的翻译性能，同时减少模型在不同领域间的翻译能力偏差。实验结果显示，与传统的随机数据选取方法相比，本文方法在口语、医学以及多领域测试集上的 BLEU 值均有显著提升。尽管如此，由于当前仅在 330 多万条单语数据上进行测试，且未能收集到涵盖资源稀缺领域相关的单语数据，在一定程度上限制了数据的多样性。此外，现有单语数据的领域分布仍然主要集中于新闻和政治领域，一定程度上制约了模型在稀缺资源领域的性能的进一步提升。

尽管相对位置编码相比绝对位置编码在性能上有所提升，但仍存在训练和预测时输入长度不一致，导致模型泛化能力下降的问题。当前，旋转位置编码等先进的编码方法能够有效解决这一问题，但受篇幅限制，本文未对这些方法进行深入探讨。未来的研究将不仅着眼于收集更多领域相关的单语数据，还将探索旋转位置编码等技术的改进，进一步验证并优化本文方法。

参 考 文 献

[1] 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.

[2] 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

[3] BROWN T B，MANN B，RYDER N，et al. Language Models are Few-Shot Learners[J]. arXiv：2005. 14165，2020.

[4] TOUVRON H，LAVRIL T，IZACARD G，et al. LLaMA：Open and Efficient Foundation Language Models[J]. arXiv：2302. 13971，2023.

[5] TOUVRON H，MARTIN L，STONE K，et al. Llama 2：Open Foundation and Fine-Tuned Chat Models[J]. arXiv：2307. 09288，2023.

[6] JAPKOWICZ N, STEPHEN S. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449.

[7] ALSOHYBE N T, DAHAN N A, BA-ALWI F M. Machine-Translation History and Evolution: Survey for Arabic-English Translations[J]. Current Journal of Applied Science and Technology, 2017, 23(4): 1-19.

[8] ARIVAZHAGAN N, BAPNA A, FIRAT O, et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges[J]. arXiv: 1907.05019, 2019.

[9] 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). Association for Computational Linguistics, 2020: 7622-7634.

[10] BRANTS T, POPAT A C, XU P, et al. Large Language Models in Machine Translation[C]// Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2007: 858-867.

[11] JAPKOWICZ N, STEPHEN S. The class imbalance problem: A systematic study[J]. Intelligent Data Analysis, 2002, 6(5): 429-449.

[12] JAMES G, DANIELA W, TREVOR H, et al. An Introduction to Statistical Learning: with Applications in Python[M]. Cham: Springer International Publishing, 2023.

[13] MOHAMMED R, RAWASHDEH J, ABDULLAH M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results[C]// 2020 11th International Conference on Information and Communication Systems. 2020: 243-248.

[14] WEISS K, KHOSHGOFTAAR T M, WANG D. A survey of transfer learning[EB/OL]. https://doi.org/10.1186/s40537-016-0043-6.

[15] KUANG J, XU G, TAO T, et al. Class-Imbalance Adversarial Transfer Learning Network for Cross-Domain Fault Diagnosis With Imbalanced Data[EB/OL]. https://doi.org/10.1109/TIM.2021.3136175.

[16] 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. Association for Computational Linguistics, 2017: 567-573.

[17] ZHANG Z, LIU S, LI M, et al. Joint Training for Neural Machine Translation Models with Monolingual Data[J]. arXiv: 1803.00353, 2018.

[18] GIBADULLIN I, VALEEV A, KHUSAINOVA A, et al. A Survey of Methods to Leverage Monolingual Data in Low-resource Neural Machine Translation[J]. arXiv: 1910.00373, 2019.

[19] SHORTEN C, KHOSHGOFTAAR T M, FURHT B. Text Data Augmentation for Deep Learning[J]. Journal of Big Data, 2021, 8(1): 101.

[20] DUAN S, ZHAO H, ZHANG D, et al. Syntax-aware Data Augmentation for Neural Machine Translation[J]. arXiv: 2004.14200, 2020.

[21] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[22] NIU X, DENKOWSKI M, CARPUAT M. Bi-Directional Neural Machine Translation with Synthetic Parallel Data[C]// Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. Association for Computational Linguistics, 2018: 84-91.

[23] ZIPF G K. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology[M]. Ravenio Books, 2016.

[24] ZHANG Y, KANG B, HOOI B, et al. Deep Long-Tailed Learning: A Survey[J]. arXiv: 2110.04596, 2021.

[25] TAN J, WANG C, LI B, et al. Equalization Loss for Long-Tailed Object Recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11659-11668.

[26] 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. Association for Computational Linguistics, 2018: 436-446.

[27] 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. Association for Computational Linguistics, 2002: 311-318.

KALZANG Gyatso, born in 1988, Ph.D candidate, is a member of CCF (No. Z0116G). His main research interests include computational linguistics and Tibetan-Chinese machine translation.

NYIMA Tashi, born in 1964, Ph.D, professor, Ph.D supervisor. His main research interests include Tibetan information technology and computational linguistics.

(责任编辑:何杨)