基于多样性数据重组增强的藏汉神经机器翻译

薛嘉怡[1,2],陈锦明[3],陈波[1,2,*],鲍薇[4,*],赵小兵[1,2]
1.中央民族大学,信息工程学院,北京,100081
2.国家语言资源监测与研究民族语言中心
3.中央民族大学,理学院,北京,100081
4.中国电子技术标准化研究院,北京,100007
{xuejy,22011151,chenbomuc}@muc.edu.cn,baowei@cesi.cn,nmzxb_cn@163.com

摘要

高资源语言的神经机器翻译虽已取得显著进展，但低资源语言面临更严重的平行数据不足的问题。为此，提出一种面向藏汉神经机器翻译的多样性数据重组增强方法（DiRec）。该方法利用大语言模型的双向语言能力，对已有藏汉平行数据进行成分重组、句型重组和风格重组三种数据重组，经过两轮质量自动筛选后得到多样性增强数据。在藏汉机器翻译的实验中，相较于基线模型，基于DiRec的模型的泛化能力指标提升4.83个百分点，BLEU提高0.55，chrF++提高0.20。最后分析了不同数据重组方式对翻译模型性能的影响。

关键词： 神经机器翻译 ； 低资源语言 ； 大语言模型 ； 多样性 ； 数据重组

A Diverse Data Recombination Augmentation for Tibetan-Chinese Neural Machine Translation

Jiayi Xue[1,2],Jinming Chen[3],Bo Chen[1,2,*],Wei Bao[4,*],Xiaobing Zhao[1,2]
1.School of Information Engineering, Minzu University of China,Beijing 100081, China
2.National Language Resource Monitoring & Research Center for Minority Languages
3.College of Science, Minzu University of China,Beijing 100081, China
4.China Electronics Standardization Institute,Beijing,100007,China

Abstract

The neural machine translation (NMT) for high-resource languages has made significant progress, but low-resource languages face a more severe problem of insufficient parallel data. To address this, this paper proposes a diversity data reorganization enhancement method (DiRec) for Tibetan-Chinese neural machine translation. This method leverages the bidirectional language ability of large language models to perform component recombination, sentence structure recombination, and style recombination on existing Tibetan-Chinese parallel data. After two rounds of automatic quality screening, diverse enhanced data are obtained. Compared to the baseline model, the DiRec-based model improves generalization ability by 4.83 percentage points, increases BLEU by 0.55, and chrF++ by 0.20. Finally, the impact of different data recombination methods on the performance of the translation model is analyzed.

Keywords: Neural Machine Translation , Low-resource Language , Large Language Model , Diversity , Data Recombination

* 通讯作者

第二十四届中国计算语言学大会论文集，第16页-第27页，济南，中国，2025年8月11日至14日。

1 引言

近年来，基于神经网络的机器翻译（Neural Machine Translation, NMT）(Sutskever et al., 2014)取得了空前进步，但对平行数据的过度依赖使其难以应用于低资源语言。此外，低资源语言通常具有独特的语法结构、词汇体系和书写规则，这些复杂的语言特性也在一定程度上限制了机器翻译模型的泛化能力。如藏汉机器翻译中，缺乏足够的高质量平行数据已成为主要瓶颈(Wang et al., 2022)，藏语的独特性进一步限制了平行数据的规模和领域覆盖度。

在低资源机器翻译中，通常采用数据增强的方法。目前典型的方法有回译、重组平行数据和平行句对挖掘等(Latief et al., 2024)。回译主要是通过翻译来创建数据的不同版本(Sennrich et al., 2016)，生成句子的质量受限于模型本身的性能及两种语言的缘近程度；重组平行数据方法通过对已有的平行数据进行启发式的替换、修改、扩展等手段，试图增加数据量和多样性(Pellicer et al., 2023)，但往往受限于启发式方式的模式有限性，无法有效提升模型的泛化能力；平行句对挖掘方法通过设计不同算法从未标注或低质量语料中识别出潜在的平行句对来扩充训练集(Chen et al., 2020)，需对不同语言及待挖掘数据的特点设计特定算法，迁移性较差。这些典型的数据增强方法虽然已经在一定程度上缓解了低资源机器翻译中数据不足的问题，但模型仍面临泛化能力不足的挑战。

为此，本文提出了一种面向藏汉低资源神经机器翻译的多样性数据重组方法（DiRec，**Di**versity Data **Rec**ombination）。该方法是一种以大模型为构造器的数据增强方法，利用大模型的双向语言能力，通过成分重组、句型重组和风格重组三种重组策略，在已有的平行语料基础上，构造成分组成、句型结构和句子风格更多样的平行数据。具体地，在给定的若干组平行数据的基础上，利用大语言模型本身的双语语言能力、双语翻译能力和任务执行能力，进行句子级语义一致的多句子间成分随意组合的数据重组，即成分重组，如将平行句A（a）的主语成分替换到平行句B（b）中的宾语成分中；进行任一句式到所有其它可能句式的重组，实现句型重组，如肯定句转疑问句、反问句、祈使句等；进行任一风格到所有可能的句子风格的重组，即风格重组，如均为正式表达的平行句重组为同为非正式表达的平行句。实验中，通过递进式添加多样性增强语料库中的数据探究不同重组策略对翻译模型性能的影响，并对来自不同重组方式得到的增强语料施加不同的权重，探索最优的数据重组比重。利用BLEU、chrF++、TER、预测间隔、输入敏感度五个指标评估翻译模型的性能。本文的贡献如下：

- 提出一种基于LLM的多样性平行数据重组的自标注方法——DiRec[1]，能有效提升藏汉神经机器翻译效果，并为解决低资源神经翻译中平行数据匮乏提供一种可能的解决方案。
- 提出三种数据重组方法，分别从句子成分的重组、句子结构的重组、句子风格的重组生成不同维度的多样性平行数据，显著提升藏汉神经机器翻译模型的泛化能力。
- 通过对不同重组方式得到的多样性平行数据进行不同权重的配比，探索翻译质量与泛化能力的协同提升的可能路径。

2 相关工作

2.1 低资源机器翻译

目前世界上现存语言超过5000种，低资源语言在其中占有重要地位，这些语言面临数据稀缺的问题，导致其机器翻译质量往往无法与富资源语言相提并论。为此，诞生了许多针对低资源机器翻译的工作。例如，Zoph等(2016)利用迁移学习将高资源语言的知识迁移到低资源翻译任务中，缓解数据稀缺带来的问题；Firat等(2016)通过联合多种低资源语言与高资源语言一起进行训练，借助语言之间的共性提升翻译效果；Nguyen等(2024)通过丰富的语言提示充分挖掘LLM的潜力，从而更适应低资源语言的特点；Mullov等(2024)将词汇从NMT模型整体的语言学习过程中分离出来，通过独立学习词汇提升模型的适用语言范围，以弥补双语数据不足带来的困境。此外，数据增强方法成为提升低资源机器翻译性能的重要手段，Jin(2024)提出一种基于语义词替换的数据增强方法，先通过替换语义词扩充语料库，再利用神经语法校对模型筛选扩充语料；Mahamud等(2023)在简易数据增强方法的基础上，使用了语义词上下文信息和词性

[1]重组数据和代码公开于：https://github.com/breezebinbin/DiRec。

标签进行词替换和增强；Shen等(2023)通过将目标端句子变换生成带噪声的伪平行语料，并设计多任务辅助学习策略强化编码器表示能力提升翻译性能；不同于已有的数据增强方法，本文的DiRec方法充分利用了LLM的理解能力和生成能力，从数据多样性的角度去重组数据，具有更广的数据分布。

2.2 藏汉机器翻译

藏语作为典型的低资源语言，与汉语的互译也成为代表性的低资源机器翻译。早期群诺等(2018)聚焦于规则与词典，构建了统计藏汉机器翻译系统，但译文可能因受限于规则泛化能力出现句式僵化问题；随着预训练语言模型的发展，Sun等(2022)提出的藏语预训练语言模型极大推动了藏文自然语言处理任务的发展，也为藏汉翻译任务注入了新动力，但在领域泛化性上仍存在一定局限。桑杰端珠等(2023)通过将藏汉双语词典与藏汉单语数据结合，利用BART风格的降噪自编码进行预训练，增强了模型在平行数据稀缺的情况下学习双语知识的能力。Shen等(2024)利用知识蒸馏思想将语言模型的先验知识迁移至翻译模型，通过目标端单语语言模型对神经机器翻译训练进行正则化的方法提升翻译效果。与此同时，各种数据增强方法被尝试应用于藏汉机器翻译(Zhuang et al., 2025)。本文针对现有的藏汉机器翻译方法在泛化性上仍明显不足的问题，提出多类型的数据重组策略，显著提升模型的泛化能力。此外，也有李林霞等关注低资源语言中平行语料的对齐质量评估2025，常润等2024则从语义层面出发，增强术语和新词的翻译效果，进一步改善了藏汉翻译的术语覆盖，为提升藏汉翻译的准确性和术语覆盖提供了有益探索。

2.3 LLM在低资源机器翻译中的应用

随着以ChatGPT(Brown et al., 2020)、ChatGLM(Zeng et al., 2024)、LLaMA(Touvron et al., 2023)为代表的LLMs通过超大规模参数训练，展现出卓越的跨语言理解与生成能力，LLM具备的上下文学习(Dong et al., 2024)、零样本/少样本泛化能力(Wei et al., 2022)以及语言无关的表征空间(Conneau et al., 2020a)等核心特性，为直接作为低资源翻译器提供了可能。然而，现有研究表明LLM在高低资源语言间的翻译性能存在显著差异。例如，Jiao等(2023)的研究指出，ChatGPT-4在高资源语言中的翻译质量接近专业翻译系统，但在低资源语言翻译任务中却显著落后；另一方面，Conneau等(2020b)的实验表明，即使对于训练数据中从未显示出现的零资源语言，LLM仍然能够生成可接受质量的翻译，这在一定程度上表明LLM可能通过预训练数据中隐含的多语言对齐模式实现了跨语言迁移(Choenni et al., 2023)。为提升LLM与低资源语言的适配效果，许多学者展开了积极探索。Zhu等人(2024)证实通过上下文学习引入跨语言示例可以有效引导LLM生成低资源翻译，其机制与Conneau等(2020b)提出的跨语言示例激活多语言表示空间相互呼应，为LLM的泛化性应用提供了方法论支撑。此外，指令微调被证明能够增强模型对低资源语言的敏感程度(Muennighoff et al., 2023)，进一步体现了任务适配性优化在缩小高低资源语言性能差距中的关键作用，也为LLM在低资源数据增强领域带来了独特的挑战和机遇。本文所提方法DiRec以LLM为数据构造器，基于多种数据重组策略，实现有效数据增强。

3 DiRec：多样性数据重组增强方法

针对低资源机器翻译中平行语料不足，模型泛化性能受限的问题，本文提出基于大模型的多样性数据重组增强方法DiRec，包含三种数据重组策略和增强数据最优选择。首先利用LLM和数据重组策略，生成重组数据，然后进行启发式增强数据选择的优化，在藏汉神经机器翻译基座模型上实现最优的数据增强效果。接下来，分别介绍方法的整体框架、三种数据重组策略、增强数据的选择优化。

3.1 DiRec整体框架

DiRec的整体框架如图1所示。首先，该方法基于已有的平行数据，以及成分重组、句型重组和风格重组三种数据重组指令，利用LLM的双向语言能力和任务执行能力，生成大量重组的平行语料。然后，对这些重组的藏汉平行语料进行两轮质量自动筛选后，形成多样性数据重组数据，极大地扩展了藏汉平行语料的数据分布。最后，以Transformer结构为神经机器翻译基座，对语料库中不同重组策略来源的平行语料进行多种组合，探索来自不同重组策略的数据对机器翻译模型性能的影响。

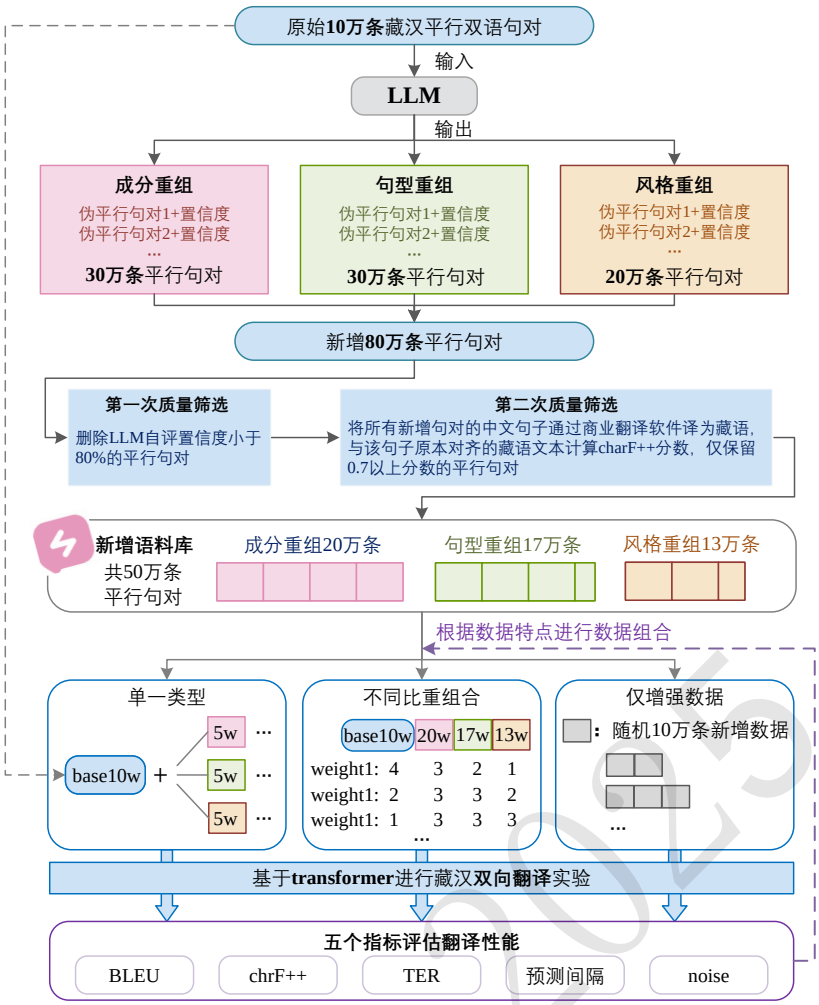

Figure 1: DiRec方法整体框架

3.2 三种数据重组策略

DiRec中包含三种数据重组策略：成分重组、句型重组和风格重组，其中成分重组是将不同句子中的成分进行两端语义一致性的交叉重组，句型重组是将原平行句子与新的句型进行同步的重组，得到不同句型的新平行句子，风格重组是将原平行句子与新的风格进行一致的重组，得到不同风格的新平行句子。数据重组均以提示大模型的方式完成，为尽可能减少对大模型生成的限制，我们对每个句子的具体变换方式不做硬性规定，由模型自行选择最合适的变换方式。大模型我们选择GPT-4o，三种数据重组的形式和提示模板如图2所示。

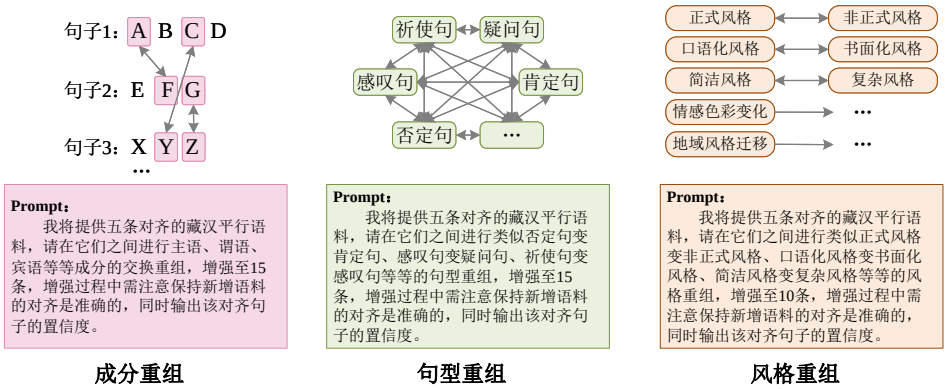

Figure 2: 三种数据重组的提示

具体以组为单位生成的示例如表1所示：

Table 1: 数据重组示例

输入LLM的原始数据	
汉文	藏文
她无法接受这样的事实。	ཁོ་མོས་དོན་དངོས་འདི་འཇུ་དང་ལེན་བྱེད་ཐབས་མེད་པ་རེད།
你见到你那三根毛的好朋友了吗?	ཁྱོད་ཀྱི་ཐོག་པོ་སན་མཁོ་ཤེ་གུ་གམ
李书记平时有什么爱好?	རྒྱུ་ཅི་ལེ་ལ་ནས་རྒྱུན་དགའ་སྤྱོགས་ཅི་ཞིག་ཡོང་དམ།
10年前，来自贵州遵义的傣族女子杨丽与嘎查蒙古族青年张景相爱。	ལོ་བཅུའི་ཧོང་ཀུའེ་ཙོག་ཁུན་དབྱིས་ཡིན་པའི་ཉའེ་རིགས་ཀྱི་ཕུད་མེད་དབྱངས་ལི་དང་ཀ་ཁའེ་མིག་རིགས་གཞོན་ནུ་ཀང་ཅིང་གཉིས་དགའ་རོགས་བསྟིགས་པ་དང་།
《报告》显示，国有文化企业资产和经营规模不断扩大。	《སྐན་ཁའེ》 ནང་མཇོན་གསལ། གཞུང་ཚོངས་རིག་གནས་ལེ་ལས་ཀྱི་རྒྱུ་རོར་དང་གཉེར་སྒྲིང་གཞི་བྱོན་རྒྱུན་ཚད་མེད་པར་ཚེ་ཐུ་ཐྱེན་ནས་
通过成分重组得到的部分增强数据	
汉文	藏文
她见到你那好朋友了吗?	མོས་ཁྱོད་ཀྱི་ཐོག་པོ་བཟང་པོ་དེ་མཇོང་སོང་ངམ།
李书记10年前有不同的爱好。	རྒྱུ་ཅི་ལེ་ལ་ལོ་བཅུའི་སྔོན་ལ་དགའ་སྤྱོགས་མེ་འདྲ་བ་ཡོད།
来自贵州的张景无法接受这个事实。	ཀུའེ་ཙོའུ་ནས་ཡོང་བའི་ཀང་ཅིན་ཀྱིས་དོན་དངོས་འདི་དང་ལེན་བྱེད་མི་ཐུབ།
你那三根毛的好朋友有多高?	ཁྱོད་ཀྱི་སྐུ་ཤང་གསུམ་ཅན་ཀྱི་ཐོག་པའི་གཟུགས་པའི་རིང་ཚད་ལ་ཅི་ཙམ་ཡོད།
有多少企业在《报告》中被提到?	ལེ་ལས་ག་ཚོད་ 《སྐན་ཁའེ་》 ནང་སྤྲོང་ཡོད།
通过句型重组得到的部分增强数据	
汉文	藏文
为什么她无法接受这样的事实?	ཁོ་མོས་དོན་དངོས་འདི་འཇུ་ཞིག་དང་ལེན་བྱེད་མེ་ཐུབ་པའི་རྒྱུ་མཚན་ཅི་ཡིན་།
你可以告诉我李书记爱好什么吗?	ཁྱོད་ཀྱིས་ང་ལ་རྒྱུ་ཅི་ལེ་ཡི་དགའ་སྤྱོགས་ཅི་ཞིག་ཡིན་པ་བཤད་ན་ཆོག་གམ།
杨丽和张景幸福地生活在一起吧！	དབྱངས་ལིས་དང་ཀང་ཅིན་གཉིས་བདེ་སྐྱིད་ལྡན་པའི་ངང་མཉམ་དུ་འཚོ་བ་རོལ།
《报告》为何显示规模增长?	《 སྐན་ཁའེ་ 》 ནང་ཅིའི་ཕྱིར་གཞི་ཁྱོན་འཕར་བ་མཇོན་པ་རེད་།
我希望她能见到更多的朋友	ངའི་རེ་བ་ཡིན་ན་མོས་ཐོག་པོ་ཤང་པོ་ཞིག་མཇོང་ཐུབ་ན་བགམ།
通过风格重组得到的部分增强数据	
汉文	藏文
尊敬的李书记，请问您的兴趣是什么?	ཀུས་པར་བོས་པའི་རྒྱུ་ཅི་ལེས།ཁྱེད་ཀྱི་དགའ་སྤྱོགས་ཅི་ཞིག་ཡིན།
哈哈，你真的见到了那个《报告》吗?	ཧ་ཧ།ཁྱོད་ཀྱིས་ངོ་མ་སྐན་ཁ་རེ་མཇོང་སོང་ངམ།
她的文学作品中的人物中总是这样坚强。	མོའི་རྩོམ་རིག་བརྩམས་ཚོས་ཁྲོད་ཀྱི་མི་སྣ་ནི་ཅམ་ཡང་འདི་འདྲའི་སྤྲ་བརྟན་འགྲུར་མེད་ཡིན།
听说你的朋友在新的报告中被提及?	གོ་མོས་ཤར་ན་ཁྱོད་ཀྱི་ཐོག་པོ་སྐན་ཁ་ཤས་བའི་ནང་དུ་སྤྲོང་ཡོད་ཅེར།
李书记对她的报告很有研究！	རྒྱུ་ཅི་ལེ་ལགས་ཀྱིས་ཁོ་མོའི་སྐན་ཁ་ཤེར་ཞིབ་པར་རྩལ་ཆེན་པོ་ཡོག་ཡོད།

3.3 增强数据的选择优化

首先，生成的藏汉平行语料经过两轮质量筛选以过滤低质量数据。第一轮筛选基于LLM自评估置信度阈值进行初步过滤，删除自评置信度小于80%的句对,这种将“评分”与“筛选”进行解耦的策略，在当前藏语资源稀缺、缺乏现成自动评估工具的条件下，是在效率与可靠性之间的一种折中选择。进行第二轮筛选时，先将所有新增平行句对的中文句子通过翻译工具[1]译为藏语，将得到的参考译文与新增语料中与该句子原本对齐的藏语文本计算chrF++分数，删除得分小于0.7的句子，以此确保增强语料的准确性。

然后，三种数据重组策略所生成的数据各有侧重，通过合理组合不同策略的语料，可以得到更高质量的平行语料，并在基座翻译模型上进行验证。我们采用了三种数据选择方式：(1)选择单一类型数据进行增强，例如，将所有基于成分重组策略生成的平行句对以5万条为单位，与原始平行数据集进行组合，每次加入一个单位，形成多组对比训练数据；(2)根据不同比重组合三种重组策略的生成数据进行增强，例如，选择更多的成分重组的数据，相对少的风格重组的数据；(3)仅使用生成的平行数据作为训练语料。在实验过程中，通过调整数据的组合方式及选择比重，来平衡训练语料的丰富性和多样性。

[1]小牛翻译：https://niutrans.com。

4 实验与分析

4.1 实验设置

4.1.1 基础平行语料

实验的基础平行语料来源西北民族大学提供的藏汉双语平行句对数据集(2022)，包含100万条藏汉平行句对，我们随机抽取了其中10万条平行句对作为本文的基础数据集。测试数据来自第17届全国机器翻译大会(CCMT 2021)提供的汉藏双语数据集，包含1379条句对，用于评估翻译模型的性能。基于基础数据，DiRec通过三类数据重组策略，新生成80万条平行句对，经过两轮质量筛选后得到成分重组类的20万条句对、句型重组类的17万条句对、风格重组类的13万条句对，共计50万条增强平行句对。预处理部分，对汉语采用jieba进行分词；对藏语采用格桑加措等(2024)提出的藏文分词方法进行分词。

4.1.2 对比实验数据

根据不同的数据选择（详见3.3节），我们设计了多组对比实验数据，具体如表2所示。其中基础数据base10w，将成分重组得到的增强数据记为CR（Component Recombination），将句型重组得到的增强数据记为MR（Mood Recombination），将风格重组得到的增强数据记为SR（Style Recombination），括号中的数字为平行数据的量。

Table 2: 不同数据组合方式的数据构成

数据组合方式	数据来源	实验序号	数据构成详情
基线数据	base10w	1	base10w
单一类型	base10w+CR (基础数据+成分重组数据)	2	base10w+CR(5w)
		3	base10w+CR(10w)
		4	base10w+CR(15w)
		5	base10w+CR(20w)
	base10w+MR (基础数据+句型重组数据)	6	base10w+MR(5w)
		7	base10w+MR(10w)
		8	base10w+MR(15w)
		9	base10w+MR(17w)
	base10w+SR (基础数据+风格重组数据)	10	base10w+SR(5w)
		11	base10w+SR(10w)
		12	base10w+SR(12w)
	base10w+CR+MR+SR	13	base10w+CR(20w)+MR(17w)+SR(13w)
不同比重组合	base10w+CR+MR+SR	14	分别设置权重为：4、3、2、1
		15	分别设置权重为：2、3、3、2
		16	分别设置权重为：1、3、3、3
仅增强数据	CR+MR+SR (仅含重组数据)	17	随机抽取20万个句对
		18	随机抽取30万个句对
		19	随机抽取40万个句对
		20	随机抽取50万个句对

4.2 基座模型

Table 3: 模型超参数设置

参数名称	值	参数名称	值	参数名称	值
编码器层数	8	解码器层数	8	注意力头数	16
优化器	adam	学习率	1	学习率衰减方法	noam
batch-size	8192	bucket-size	262144	warmup-steps	4000
梯度累计次数	4	隐层维度	1024	前馈网络维度	4096

我们基于自注意力机制的经典Transformer架构构建神经机器翻译模型，意在突出数据重组策略本身的增益。实验依托OpenNMT-py实现(Klein et al., 2017)。关键参数设置如表3所示，所有实验均采用以上参数。实验使用Ubuntu20.04操作系统，在Nvidia Tesla V100S PCIE 32GB完成训练，平均每次实验花费26个GPU小时。

4.3 评估指标

Table 4: 评价指标详情

指标名称	衡量维度	详情
BLEU(Papineni et al., 2002)	流畅度	计算翻译结果与参考翻译之间的n-gram重合度，本文取n-gram为1-3的综合得分
chrF++(Popović, 2017)	准确性	计算字符级别的n-gram匹配情况
TER(Snover et al., 2006)	编辑成本	计算模型输出经过多少次变换才可与目标输出一致，反映模型输出的译后编辑成本
预测间隔(Li et al., 2019)	泛化能力	计算第一候选句子和第二候选句子之间的概率差异，反映模型对未见数据的适应能力
输入敏感度	鲁棒性	受到Li等(2019)启发，引入noise1通过随机交换测试集中包含六个以上token的句子中两个token的位置来模拟翻译中的语法或词序变化；noise2在noise1的基础上，进一步随机删除一个token，将noise1、noise2输入模型；通过计算模型在noise1和无噪声输入上的BLEU分数变化度量模型的顺序鲁棒性，通过计算模型在noise1和noise2上的chrF++分数变化度量模型的字符鲁棒性

为全面评估机器翻译的效果，本文采用多个评价指标，涵盖流畅度、准确性、编辑成本、泛化能力、模型鲁棒性共5个维度。各指标的详细描述见表4。

本文进行了汉文译为藏文、藏文译为汉文两个方向的实验。每次实验训练均进行10万步，每进行1万步保存一次模型检查点，最终取该实验所有模型检查点中综合性能最优模型的数据记为实验数据。具体的汉藏实验结果见表5，藏汉实验结果见表6。在这两张表格中，表头指标后的箭头标识了每个指标的最优表现方向，并加粗了每个指标的最优数据。

Table 5: 汉藏翻译方向实验结果

数据构成		实验序号	BLEU↑	noise1 BLEU↓	chrF++↑	noise2 chrF++↓	TER↓	预测间隔↑
base10w		1	47.25	4.06	59.54	4.55	51.56	35.27%
单一类型								
base10w+CR	5w	2	46.61	4.24	58.95	4.85	52.03	39.23%
	10w	3	45.48	4.63	58.02	4.15	52.98	40.61%
	15w	4	45.52	5.27	57.74	4.63	52.63	41.02%
	20w	5	45.16	4.34	57.68	4.71	52.38	40.40%
base10w+MR	5w	6	**47.68**	4.58	**59.75**	4.56	**51.27**	37.28%
	10w	7	47.13	4.25	59.25	4.54	52.13	39.50%
	15w	8	45.72	**3.62**	58.01	4.58	52.58	39.72%
	17w	9	46.84	4.46	58.81	4.00	51.76	39.87%
base10w+SR	5w	10	45.03	4.25	57.38	4.34	53.51	39.17%
	10w	11	45.05	4.21	56.95	**3.71**	53.70	40.36%
	12w	12	44.14	3.68	56.32	3.91	54.82	**41.33%**
base10w+CR+MR+SR	60w	13	44.68	5.09	56.79	4.30	53.35	37.17%
不同比重组合								
base10w+CR+MR+SR	4:3:2:1	14	**46.49**	5.56	**58.55**	**4.42**	**51.70**	34.32%
	2:3:3:2	15	44.87	4.49	56.85	5.15	53.41	37.19%
	1:3:3:3	16	43.28	**4.14**	55.99	4.50	54.21	**38.32%**
仅增强数据								
CR+MR+SR	20w	17	18.72	1.18	31.47	**2.09**	76.45	**75.36%**
	30w	18	22.48	**0.96**	36.17	3.22	72.44	68.74%
	40w	19	24.90	1.51	38.97	3.15	69.86	62.71%
	50w	20	**25.91**	1.78	**39.38**	3.29	**69.38**	60.35%

Table 6: 藏汉翻译方向实验结果

数据构成		实验序号	BLEU↑	noise1 BLEU↓	chrF++↑	noise2 chrF++↓	TER↓	预测间隔↑
base10w		21	42.99	4.49	39.19	2.96	54.75	55.38%
单一类型								
base10w+CR	5w	22	43.21	4.84	39.20	**2.90**	55.82	58.36%
	10w	23	43.56	4.39	39.37	3.16	55.96	56.98%
	15w	24	44.07	4.55	**39.73**	3.37	56.64	57.90%
	20w	25	**44.25**	4.54	39.66	3.49	56.61	58.52%
base10w+MR	5w	26	43.40	5.03	39.04	2.93	54.72	55.45%
	10w	27	43.18	4.12	39.09	3.16	55.79	58.00%
	15w	28	43.29	4.21	39.01	3.22	55.80	58.68%
	17w	29	43.18	**3.93**	38.99	3.50	55.83	58.74%
base10w+SR	5w	30	43.33	4.36	39.30	3.51	54.94	57.01%
	10w	31	43.54	4.57	39.39	3.07	**55.22**	**60.21%**
	12w	32	43.23	4.49	39.03	3.58	55.25	57.25%
base10w+ CR+MR+SR	60w	33	**44.25**	4.71	**39.66**	3.75	56.61	55.11%
不同比重组合								
base10w+ CR+MR+SR	4:3:2:1	34	**45.70**	5.21	**40.84**	**3.18**	**54.15**	52.76%
	2:3:3:2	35	44.49	4.75	39.80	3.60	55.93	**55.16%**
	1:3:3:3	36	43.07	**4.37**	38.35	3.80	59.28	54.72%
仅增强数据								
CR+MR+SR	20w	37	19.40	**1.48**	17.47	1.38	89.64	**83.78%**
	30w	38	24.19	2.68	21.68	1.60	84.95	77.65%
	40w	39	25.77	3.07	23.71	1.43	**84.46**	73.47%
	50w	40	**27.86**	3.04	**25.22**	2.20	80.76	71.47%

分析表5、表6数据可知：

（1）多样性重组增强语料可以显著提高翻译模型的泛化能力。在汉文和藏文的双向翻译实验中，当同时使用基础数据与单一类型的增强数据训练模型时，汉藏翻译实验12即SR(12w)较基线实验1预测间隔分数提升6.06%，汉藏翻译实验31即SR(10w)较基线实验21预测间隔分数提升4.83%。当仅使用增强数据训练模型时，即实验17-20及实验37-40，模型泛化能力的提升更加显著，BLEU分数、chrF++分数随着数据量的增大呈稳步上升趋势，同时编辑成本也随之降低，泛化能力与模型鲁棒性虽逐渐下降，却仍显著优于基线实验1，例如汉藏翻译实验17较实验1预测间隔分数提升40.09%，藏汉翻译实验37较实验21预测间隔分数提升28.41%。再结合所有实验数据来看，可以得到一个明确结论：不论是单独训练增强数据，还是将增强数据与原始数据结合，均能够提高翻译模型的泛化能力。出现这一现象的原因可归因于DiRec方法通过利用LLM，通过三种重组方式生成了不同成分、句型和风格的藏汉平行数据，这种数据增强方法不仅增加了数据量，还使数据覆盖了更多的应用场景和表达方式，模型通过学习这些被增强的数据，能够学习到更加多样化和复杂的语言规律，从而提升其泛化能力。具体来说，成分重组通过改变句子的组成成分，令模型学习到更丰富的语法结构；句型重组通过改变句子的句型结构，令模型对语言中不同表达方式的理解更加深刻；风格重组通过调整句子的风格、语气等，拓宽了模型对不同语言风格的适应能力。得益于DiRec的数据增强策略，数据分布通过这些重组方式得到了扩张，使模型通过学习这些数据接触到更加丰富的语言输入，从而实现泛化能力的显著提升。

（2）藏汉翻译中，三种数据重组策略所生成的数据都具有很好的增强效果。在实验31即SR(10w)中，在BLEU提高0.55、chrF++提高0.20、预测间隔提高4.83%的同时维持编辑成本仅降低0.47，noise1与noise2分别升高0.08、0.11。在其它递进式加入增强语料的实验22-32中，其BLEU分数与预测间隔分数均优于基线实验21；chrF++分数在成分重组实验与风格重组实验上均实现了提高，但在句型重组实验中有所下降；TER指标与模型鲁棒性相较于基线实验1均稍显落后。这显著证明了在藏汉翻译中，使用DiRec方法进行数据增强是十分有效的。因此可以得到一个明确结论：为了提升以低资源语言为源语言时的翻译模型在目标语言上的泛化能力、准确性及流畅度，将LLM产生的伪平行数据与原始数据结合是有效的方法。

（3）汉藏翻译中，句型重组策略所生成的数据具有更好的增强效果。在汉藏翻译实验中，仅有实验6即MR(5w)的结果较基线实验1有所提升，表现为BLEU提高0.43、chrF++提高0.21、预测间隔提高2.01%、TER降低0.29。而其它加入了增强语料的实验结果的表现均弱于基线模

型，这表明，当翻译任务的目标语言是低资源语言时，通过LLM增强对齐数据可能是有效的方法，但需要进行多次实验以筛选出可靠数据。究其原因，是由于LLM在同时变换藏文和汉文时，使用的是基于汉文学习得到的语义空间，而藏文与此空间契合程度较低，导致增强语料的对齐效果并不理想。即LLM虽然被要求生成对齐的藏汉平行语料，但其生成的汉文质量更好，而生成的藏文质量较差。换言之，LLM可能对藏文进行了错误的变换，导致增强了错误的藏文数据，却对齐了正确的汉文句子。这使得以汉文为目标语言的模型通过学习增强语料，获得了即使面对错误的藏文句子，也依然能够输出正确的汉文翻译的能力。然而，当翻译方向相反时，模型性能则会被错误的目标藏文句子影响，导致翻译质量下降。

（4）增强数据的选择优化可为不同的翻译任务提供可靠的组合方式。例如在藏汉实验34上，weight(4:3:2:1)出现了最优的BLEU分数45.70和chrF++分数40.84，较基线模型分别提高了2.71和1.65，还伴随着编辑成本TER降低了0.60，而预测间隔却降低了2.62%，noise1、noise2分别升高0.72、0.22。这表明对原始数据与成分重组、句型重组、风格重组三种增强数据按4：3：2：1的权重进行训练，更适合例如文件这类对准确性、流畅度要求严格，同时希望节约编辑成本而不看重泛化能力及模型鲁棒性的任务。而weight(2:3:3:2)的表现则更加均衡，BLEU分数和chrF++分数较基线模型分别提高了2.50和0.61，TER降低1.18，预测间隔仅降低了0.22%，noise1、noise2分别升高0.26、0.64。这表明对原始数据与成分重组、句型重组、风格重组三种增强数据按2：3：3：2的权重进行训练，更适用于例如日常生活这类对各项指标要求均衡的任务。由此可得到一条明确结论：针对增强语料的特点，结合其具体应用场景对翻译模型进行有针对性的选择与优化，是提升翻译效果的关键。

5 总结

本文提出了DiRec 方法，利用大语言模型的双向语言能力，对现有藏汉平行语料进行成分重组、句型重组和风格重组三种方式的数据增强，并通过两轮质量筛选构建高质量增强语料库，旨在缓解低资源语言翻译中平行数据稀缺和分布不足的问题。实验结果表明，增强语料显著提升了翻译模型的泛化能力：在藏汉翻译任务中，三种重组策略均展现出良好的增强效果；在汉藏翻译中，句型重组策略生成的数据表现出更优的增强效果。此外，针对增强数据的选择与优化，可为不同翻译任务提供更具针对性的组合方案。综上所述，DiRec 方法为低资源语言的机器翻译提供了一种有效的数据增强策略。未来工作可以进一步探索如何优化生成数据的质量和增强策略。。

致谢

感谢所有匿名审稿人的宝贵意见。本项研究成果受国家社会科学基金重大项目（22&ZD035）资助。

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu. Parallel sentence mining by constrained decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1672–1678, 2020.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641, 2023.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020a.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, 2020b.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, 2024.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, 2016.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

Bo Jin. Neural machine translation based on semantic word replacement. In *Proceedings of the 2024 International Conference on Generative Artificial Intelligence and Information Security*, page 106–112, 2024.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, 2017.

Andi Djalal Latief, Asril Jarin, Yaniasih Yaniasih, Dian Isnaeni Nurul Afra, Elvira Nurfadhilah, Siska Pebiana, Nuraisa Novia Hidayati, and Radhiyatul Fajri. Latest research in data augmentation for low resource language text translation: A review. In *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pages 185–190, 2024.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5689–5695, 2019.

Mosleh Mahamud, Zed Lee, and Isak Samsten. Distributional data augmentation methods for low resource language. *arXiv preprint arXiv:2309.04862*, 2023.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, 2023.

Carlos Mullov, Quan Pham, and Alexander Waibel. Decoupled vocabulary learning enables zero-shot translation from unseen languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6693–6709, 2024.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132: 109803, 2023.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, 2017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, 2016.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006.

Yuan Sun, Sisi Liu, Zhengcuo Dan, and Xiaobing Zhao. Question generation based on grammar knowledge and fine-grained classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6457–6467, 2022.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, 2022.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, 2024.

Wenhao Zhuang, Dawa Cairen, Pengmao Cairang, and Yuan Sun. Ccmt2024 tibetan-chinese machine translation evaluation technical report. In *Machine Translation*, pages 119–127, 2025.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, 2016.

常润, 陈波, and 赵小兵. 基于语义信息扩充的汉藏短语翻译语料sect. 中国科学数据, 9(4):1–7, 2024.

李林霞, 陈波, 周毛克, and 赵小兵. 面向低资源语言机器翻译的平行语料句对齐评分. 数据分析与知识发现, pages 121–132, 2025.

格桑加措, 阿卜杜热西提·热合曼, 尼玛扎西, 面加, 肖桐, and 朱靖波. Bilstm和crf结合的藏文分词方法研究. 中央民族大学学报（自然科学版）, 33(3):40–46, 2024.

桑杰端珠and 才让加. 基于词典注入的藏汉机器翻译模型预训练方法. 中文信息学报, 37(8):43, 2023.

申影利, 周毛克, and 赵小兵. 基于多任务学习的民汉神经机器翻译数据增强方法. 中文信息学报, 37(2):97, 2023.

申影利and 赵小兵. 语言模型蒸馏的低资源神经机器翻译方法. 计算机工程与科学, 46(04):743, 2024.

群诺, 尼玛扎西, 完么扎西, and 嘎玛扎西. 基于统计的汉藏机器翻译系统关键技术研究与实现. 高原科学研究, 2(2):8, 2018.

西北民族大学. 藏汉双语平行句对数据集(v1), 2022. 国家基础学科公共科学数据中心发布.