



文献 CSTR:
32001.14.11-6035.csd.2025.0090.zh



文献 DOI:
10.11922/11-6035.csd.2025.0090.zh
数据 DOI:
10.57760/sciencedb.25100

文献分类: 计算机科学技术

收稿日期: 2025-05-14
录用日期: 2025-11-14
发表日期: 2026-02-26

XBMU-MC: 多语言平行语料库

严琦栋^{1,2}, 马宁^{1,2*}, 巴桑珠扎³, 白玛曲扎³, 艾科拜·依米提³,

麦迪那木·吾斯曼³, 木巴热克·阿布力克木³, 苏日娜⁴, 阿力娅⁴

1. 西北民族大学, 语言与文化计算教育部重点实验室, 兰州 730030
2. 西北民族大学, 甘肃省民族语言文化智能信息处理重点实验室, 兰州 730030
3. 西北民族大学, 数学与计算机科学学院, 兰州 730030
4. 西北民族大学, 中国语言文学学部, 兰州 730030

摘要: 多语言平行语料库是自然语言处理领域突破语言壁垒、提升跨语言任务效率和质量的基础资源。针对当前汉藏、汉维、汉蒙等低资源语言平行语料相对稀缺的现状, 本研究构建了一个面向多语言机器翻译和跨语言信息检索任务的平行语料库——Northwest Minzu University-Multilingual Corpus (XBMU-MC)。原始语料包括人工构造特定双语文本和网络爬取公开多语言数据。人工构造语料由研究人员围绕文化、科技、社会等领域撰写汉语文本, 并通过机器翻译与人工校对获得目标语言译文; 网络爬取语料则从主流藏文、蒙文和维文新闻网站抓取原始文本, 扩大语料规模和领域覆盖范围。之后采用数据增强技术对现有语料进行扩展, 并对数据进行严格筛选, 确保每对平行语料在源语言和目标语言之间的对齐一致性与翻译准确性。最终筛选出 21,579 条高质量样本, 以 JSON 格式存储, 包括指令、输入和输出三个属性。最后, 基于 GLM4-9B、Qwen3-8B 和 Deepseek-R1-8B 模型对 XBMU-MC 数据集与开源 MMDS 数据集进行多语言翻译对比实验, 在 BLEU 得分上分别平均提升 21.16、20.51 和 15.37 个百分点。本数据集能在一定程度上缓解低资源语言平行语料不足的问题, 可有效支持模型训练与多语言翻译任务, 也可作为多语种大语言模型在跨语言对齐、指令微调及民族语言智能信息处理应用的语料基础, 助力相关任务性能提升。

关键词: 汉藏; 汉蒙; 汉维; 平行语料库; 数据增强

数据库(集)基本信息简介

数据库(集)名称	XBMU-MC: 多语言平行语料库
数据通信作者	马宁 (maning@xbmu.edu.cn)
数据作者	严琦栋、巴桑珠扎、白玛曲扎、艾科拜·依米提、麦迪那木·吾斯曼、木巴热克·阿布力克木、苏日娜、阿力娅、马宁
数据量	56.22 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.25100

* 论文通信作者

马宁: maning@xbmu.edu.cn

基金项目	国家自然科学基金项目（62466052）；甘肃省中央引导地方科技发展资金项目（25ZYJA034）；西北民族大学中央高校基本科研业务费专项资金项目（31920250008）
数据库（集）组成	数据集包括 21,579 条 json 格式的数据，每一条 json 对象包含三个属性，分别为 instruction、input 和 output。

引言

大语言模型（Large Language Model, LLM）在对话、问答等多个领域有广泛应用^[1]。多语言平行语料库是提高机器翻译^[2]、跨语言信息检索^[3]等任务效率和质量的核心基础资源。汉语、英语等高资源语言的平行语料库建设已相对成熟，但低资源语言如藏语、蒙古语、维吾尔语的平行语料库建设仍存在数据稀缺、质量良莠不齐、领域覆盖不足等问题，严重制约了跨语言信息处理技术的提升与应用。这一问题的产生与数据采集难度高、标注成本高昂有关。

目前已有一些研究构建了相关数据集，从一定程度上丰富了低资源语言可用资源。藏语知识图谱数据集 TiKG-30K^[4]聚焦实体关系三元组，包括 146,679 个三元组，30,986 个实体和 641 种关系，可用于知识图谱的表示学习及下游任务；中-蒙-藏-维文多文档摘要数据集 Multi-language Multi-document Summarization dataset (MMDS)^[5]包括 1,044 个新闻事件、6,234 篇新闻文章，4 种语言间高度对齐，适用于跨语言理解与摘要生成研究；TibNER 藏文命名实体识别数据集^[6]收录了 20,096 句标注语料，覆盖新闻、社会、科技等多领域；TN-SUM 藏文文本摘要数据集^[7]收集了 2 万篇藏文新闻，内容丰富；蒙汉机器翻译校正数据集^[8]基于 CCMT 2021 语料，构建 5 万句对高质量平行语料，可用于蒙汉机器翻译建模与评估；汉维可比语料数据集^[9]从网络采集新闻与政府公文，构建 5,000 篇章的汉维对应文本，适用于可比语料挖掘与跨语言处理研究；常润等构建的基于语义信息扩充的汉藏短语翻译语料 SECT 数据集^[10]，收录 7,000 条汉藏短语翻译对，针对短语翻译中语义信息不足问题，创新性地引入汉语释义与例句进行增强，为短文本翻译质量提升提供了新思路；庞仙等人发布的基于新闻文本的汉藏新词语数据集^[11]，面向汉藏新闻语料，抽取并整理出 3,388 条汉语新词和 1,417 条藏语对译新词，对词汇层级的动态演化建模具有重要价值。与国内语料库工作相比，国际上也有多个主流的多语言语料库为低资源语言的研究提供了支持。例如，FLORES-101^[12]是一个多语言评估数据集，覆盖了 101 种语言，适用于机器翻译和跨语言任务的评估，其标注机制较为严格，旨在提升低资源语言的机器翻译质量；Tatoeba^[13]是一个开放的多语言语料库，涵盖了多种语言对，尤其适用于语义表示和句法分析；OPUS^[14]提供了广泛的多语言平行语料，涵盖了数十种语言，适用于机器翻译、跨语言信息检索等任务。

尽管上述数据集在数据规模、领域覆盖和质量控制方面取得了明显进展，但从实际应用角度来看，面向低资源语言的多语言语料库规模依旧偏低，仍无法支持生成式人工智能时代对多语言数据资源的高度需求。相较于同类开源数据集，本研究构建的 XBMU-MC 平行语料库通过引入数据增强与 MQM 评估框架，实现了汉藏、汉蒙、汉维语料的高质量对齐，并在主流大语言模型多语种微调任务中表现出良好性能增益，可作为低资源语言、跨语言研究的可靠数据基础。

1 数据采集和处理方法

1.1 数据采集

本研究构建的汉藏、汉维、汉蒙多语言平行语料库，数据来源包括人工构造语料和网络爬取语料两部分。采集流程如图 1 所示。

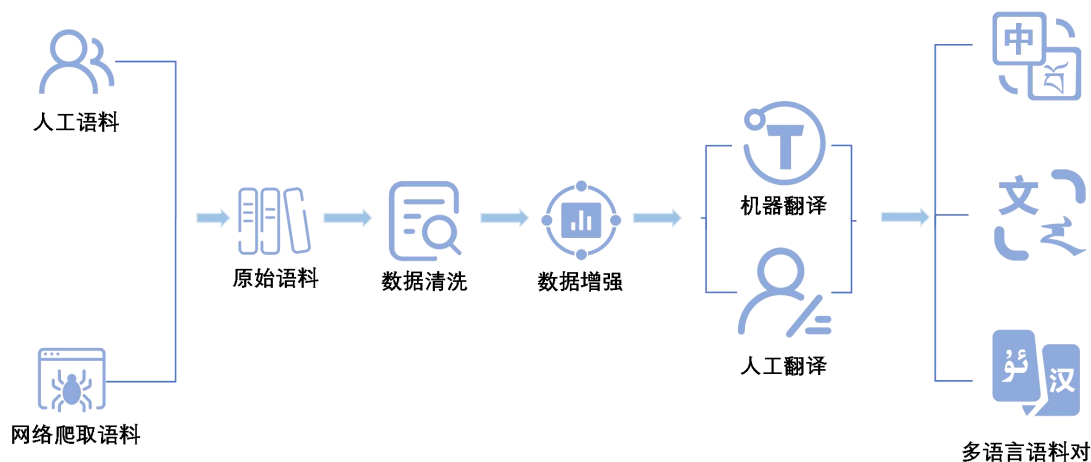


图 1 数据采集流程图

Figure 1 Data collection flowchart

1.1.1 人工构造语料

人工构造语料采用定向撰写、机器翻译与人工校对方法。首先由研究人员围绕文化、科技、社会、政治等领域定向撰写汉语源文本，再利用多语言机器翻译系统生成初步译文，最后由双语专家对译文进行严格校对，确保跨语言语义的准确对齐，形成高质量的平行语料。各领域语料的分布情况如图 2 所示。

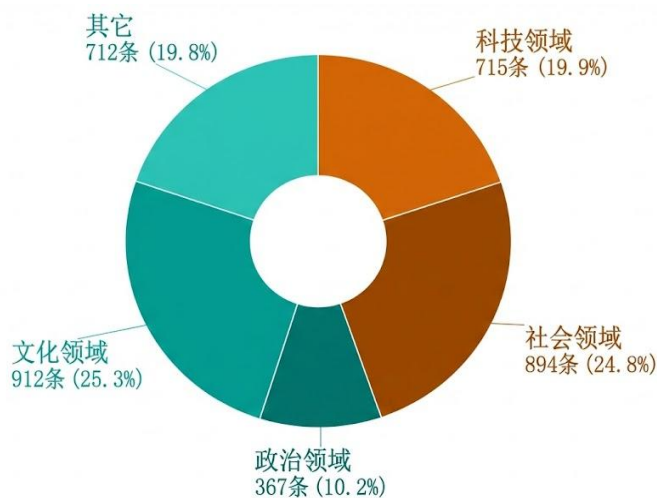


图 2 人工构造语料领域分布图

Figure 2 Domain distribution of the manually constructed corpus

1.1.2 网络爬取语料

网络爬取的生语料（Raw Corpus）来源于主流藏文、蒙古文、维吾尔语新闻网站，共 6 个，包括人民网（藏文版）、新华网（藏文版）、蒙古文文库等，具体如表 1 所示。

表 1 语料来源网站

Table 1 Website sources of the corpus

序号	网站	URL
1	人民网（藏文版）	http://tibet.people.com.cn/
2	新华网（藏文版）	http://xizang.news.cn/
3	青海藏语网络广播电视台	http://www.qhtb.cn/
4	新华网（蒙文版）	http://www.nmg.xinhuanet.com/mg/
5	蒙古文文库	https://doc.onon.cn/
6	中国维吾尔语广播网	http://www.uycnr.com/

本研究从上述公开新闻网站或政府门户定向采集多语言平行语料。爬虫程序严格遵守各网站的 Robots 排除协议，设置了合理请求频率，以降低对目标服务器的访问压力。所有采集内容均为互联网公开资源。所有原始文本都经过了严格的脱敏处理，已剔除所有可能涉及的个人身份信息与敏感内容。

1.2 数据处理

1.2.1 数据清洗与格式化

数据清洗的目的是提升数据质量，确保数据适用于后续模型训练。对采集到的原始数据进行筛查，剔除缺失关键字段或数据长度过短的样本。数据中可能包含具有潜在风险的敏感词汇或不符合社会伦理与文化规范的内容，也会被严格过滤。

1.2.2 数据增强

数据增强用于增加数据集规模和多样性，提升模型应对各种语言变化和复杂情况的泛化能力^[15]。本研究通过同义词替换和语法重组技术，对现有汉语语料的表达方式进行变换，生成语义一致但形式不同的多样化样本。将汉语文本翻译成其他语言，并通过回译（Back Translation）技术将其转回汉语，进一步增强语料多样性。表 2 展示了数据增强的具体示例，其中数据增强主要应用于汉语的部分。

表 2 数据增强示例

Table 2 Examples of data enhancement

类型	Instruction	Input	Output
原始语料	请你将以下藏语语句翻译为汉语语句:\n	板块指数的低迷表现，在带动大部分个股股价下跌的	ཐག་ལེན་གྱི་སྒྲིན་གྲངས་དམའ་བའི་མཚན་རྒྱུ་ལས་མ་ ཀླང་གི་ཀླང་དངུལ་མང་ཆེ་བའི་ཟིན་གོང་ཆག་བྱུང་སྐྱུལ་

1.2.3 数据翻译与优化

2 数据样本描述

表3 数据集中各语言对样本数量

语言对	样本数量
汉藏数据集	17,727 条
汉维数据集	1,497 条
汉蒙数据集	2,355 条

表 4 数据集样本示例

字段名称	字段说明	数据示例
“instruction”:	描述任务类型的指令提示词	请你将以下藏语语句翻译为汉语:\n
“input”:	模型的输入文本	权威人士 16 日向中国证券报记者透露, 最新的《企业年金基金监督管理办法》有望于近期正式出台。

字段名称	字段说明	数据示例
“output”:	模型的预期输出文本	ཅས་16ཉིན་དབང་གཤམ་མེ་ལྷན་གྱི་དངུལ་འཛིན་ཚགས་པར་ཐོག་གསར་འགོད་པར་《ཁོ་ལས་ཀྱི་ལོ་ལོ་ལོ་ལོ་ ཐེངས་ཚེ་ལྷ་སྐུ་དྲིད་མ་བྱ་ཐབས》 གསར་ཐོས་ཉེ་ཆར་དངོས་སུ་འདོན་རེ་ཡོད་ཅེས་བརྗོད་ཡོད།

3 数据质量控制和评估

3.1 数据质量控制

本研究根据李林霞等人^[16]提出的基于给定候选余弦与邻近候选余弦的对齐评分方法进行数据质量控制，并结合具体实验需求对评分机制进行了调整与简化，量化翻译后文本与原始文本之间的相似度，保持文本的语义一致性和逻辑连贯性。具体步骤如下：

(1) 多语言句嵌入模型采用开源工具 LASER，对翻译前后文本进行编码，将其映射为统一语义空间，固定维度向量，实现多语言间的语义对齐。

(2) 针对每一对原文与翻译文本，计算对应向量的余弦相似度，作为“给定候选相似度分数”。

(3) 为进一步评估翻译文本的语义合理性与文内连贯性，引入“邻近候选相似度”。对每一句翻译文本，计算它与同一语料中其余翻译句子的余弦相似度，选取其中相似度最高的 k 个邻近句子（本实验中 $k=4$ ），计算其与当前翻译句的平均余弦相似度。

(4) 基于给定候选相似度与邻近候选相似度，计算每一句翻译的最终对齐评分，绘制对齐评分分布图（如图 3 所示，以藏语为例）。若一条翻译句对原句相似度较高，但与其他句子相似度较低，则其对齐评分更高。

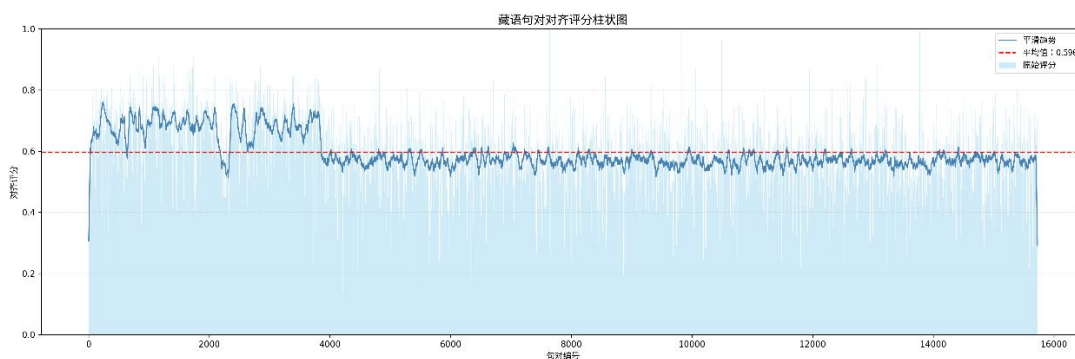


图 3 藏语句对对齐评分柱状图

Figure 3 Bar Chart of Tibetan sentence alignment scores

候选余弦相似度计算公式如下：

$$sim_{given}(i) = \cos(\vec{e}_i, \vec{e}_{i+1}) \quad (1)$$

其中， $\vec{e}_i$ 表示第 i 个翻译文本的嵌入向量。

邻近候选余弦相似度计算公式如下：

$$sim_{others}(i) = \{\cos(\vec{e}_i, \vec{e}_j) | j \neq i, j = i+1\} \quad (2)$$

$$topk_i = TopK_k(sim_{others}(i)) \quad (3)$$

$$avg_{topk}(i) = \frac{1}{k} \sum_{j=1}^k topk_i(j) \quad (4)$$

其中, $\vec{e}_j$ 表示同一语料中除第 i 句和其对应原文外的其他句子的嵌入向量; $TopK_k(\cdot)$ 表示从 $sim_{others}(i)$ 中选取相似度最高的前 k 个值 (本实验中 $k=4$); $topk_i(j)$ 表示第 j 个最高邻近相似度值。 $avg_{topk}(i)$ 表示公式核心输出结果, 指第 i 条翻译句对应 Top-k 邻近候选余弦相似度的算术平均值, 是衡量该翻译句与语料内其他句子平均语义相似程度的核心指标。

归一化对齐评分计算公式如下:

$$Score(i) = \frac{sim_{given}(i)}{avg_{topk}(i) + \epsilon} \quad (5)$$

其中, ϵ 为平滑因子。

本研究基于上述对齐评分设定语言对筛选阈值, 用以剔除低质量翻译样本。在综合考虑评分分布的平均值、平滑趋势与实验中 BLEU 得分变化的基础上, 分别为藏语、蒙古语与维吾尔语设置了如下阈值: 藏语 0.35, 蒙古语 0.20, 维吾尔语 0.20。得分过低的样本被视为可能存在翻译偏差、语义漂移或冗余问题, 将其从数据集中剔除。对齐评分再经筛选后的数据统计如表 5 所示。

表 5 对齐评分筛选后数据统计表

Table 5 Data statistics after alignment scoring and filtering

语种	原始句对数	筛选后句对数	剔除样本数	剔除原因统计		
				对齐错误占比	语义歧义	其他
藏语	18100	17727	373	41.5%	52.9%	5.6%
蒙古语	2501	2355	146	47.9%	43.1%	9.0%
维吾尔语	2589	1497	1092	44.3%	51.9%	3.8%

从统计结果可以看出, 汉维语料剔除率最高, 主要原因是原始文本中存在较多对齐不准确和语义歧义的情况, 这与语料来源分散、语言标注者数量有限密切相关。相比之下, 藏语和蒙古语语料虽然也存在一定的错误比例, 但总体质量相对较高。

3.2 数据评估

3.2.1 人工评估

本研究邀请精通藏语、蒙古语和维吾尔语的教师和学生分别对翻译数据进行打分, 评估文本对齐一致性、准确性和流畅性^[17]。对齐一致性评估多语言数据在源语言与目标语言之间对齐准确程度; 准确性评估翻译内容有无语法错误、拼写错误; 流畅性评估翻译文本是否存在语法问题, 阅读是否通顺流利。为确保翻译质量的评估更加严谨和规范, 本研究对数据集采用基于 MQM (Multidimensional Quality Metrics) 计分卡的方法^[18], 结合翻译错误的类型与严重程度进行量化评估。翻译错误被划分为四个等级, 分别为无错误、轻微错误、主要错误和严重错误, 对应分值为 0 分、1 分、5 分和 10 分。设定扣分阈值为 10 分, 若某条翻译得分超过 10 分, 则该翻译将被标记为需要人工校正和改写。对于得分极高的翻译样本则采取舍弃处理。评估打分规则如表 6 所示。

表 6 评估打分规则表

Table 6 Evaluation scoring guidelines

评估指标	无错误	轻微错误	主要错误	严重错误
对齐一致性	0-1	2-4	5-7	8-10
准确性	0-1	2-4	5-7	8-10
流畅性	0-1	2-4	5-7	8-10

人工评分结果如表 7 所示。

表 7 人工评估结果统计

Table 7 Summary of human evaluation results

语言对	对齐性得分	准确性得分	流畅性得分	合格样本占比	需人工修改样本占比
汉藏	5.7	5.5	5.2	87.5%	12.5%
汉蒙	4.4	4.0	3.5	94.0%	6.0%
汉维	3.8	3.5	3.0	83.0%	17.0%

通过观察表 7，现阶段多语言翻译的流畅性得分普遍偏低，显示出在翻译自然流畅度方面的不足，这是本研究后续工作优化的重点。

3.2.2 实验评估

为评测 XBMU-MC 数据集的质量，本研究基于大语言模型（GLM4-9B 模型、Qwen3-8B 模型和 Deepseek-R1-8B 模型），使用 XBMU-MC 数据集和开源 MMDS 数据集进行多语言机器翻译对比实验。三种基座模型均为 Transformer 架构，具备良好的跨语言建模与生成能力。数据集按 8:1:1 划分为训练集、验证集和测试集。实验以多语言文本为输入，汉语译文为输出。评估采用 BLEU-4 指标，评估模型生成译文与参考译文之间的相似度与一致性。实验使用的超参数设置如表 8 所示。

表 8 实验超参数设置

Table 8 Hyperparameter settings for experiments

Parameter	Value
batch_size	4
Epochs	3
learning_rate	5e-5
warmup_steps	3
max_input_length	2048
max_target_length	1024

3.2.3 实验结果及分析

实验结果如表 9 所示。

通过对实验结果进行分析，三个模型在 XBMU-MC 数据集上的得分分别比 MMDS 数据集平均

提升 21.16（GLM4-9B）、20.51（Qwen3-8B）和 15.37 个百分点（Deepseek-R1-8B），说明 XBMU-MC 数据集能更好地支持多语言翻译任务。

表 9 大语言模型多语言翻译 BLEU 得分对比（单位：%）

Table 9 Comparison of BLEU scores for multilingual translation of large language models (unit: %)

模型	藏语		蒙古语		维吾尔语	
Models	MMDS	XBMU-MC	MMDS	XBMU-MC	MMDS	XBMU-MC
GLM4-9B	11.25	28.92	2.93	19.65	24.38	53.48
Qwen3-8B	4.87	21.46	3.50	20.94	17.40	44.90
Deepseek-R1-8B	5.32	17.43	2.03	12.03	19.51	43.51

需要说明的是，维吾尔语数据集在各模型中的 BLEU 得分显著高于藏语和蒙古语，且在 XBMU-MC 和 MMDS 两个数据集上表现一致。经分析，这一结果是由于维吾尔语语料主要集中在新闻、政务等通用领域，这些内容与大模型预训练语料在主题上具有较高的重合度，从而提升了模型的翻译表现；同时，维吾尔语样本总量相对较小，这也导致模型在训练过程中出现了一定程度的过拟合，从而在 BLEU 指标上取得了较高分数。

此外，本数据集中各语言对样本分布存在不均衡现象，汉藏、汉蒙、汉维语料分别占 82.1%、10.9%和 6.9%。这种分布差异直接影响了模型的学习效果，汉藏语料充足，模型性能稳定；汉维语料虽少但领域集中，导致 BLEU 得分偏高但泛化能力有限。数据不均衡在一定程度上制约了模型的跨语言泛化性能，后续研究需进一步优化各语言对的数据分布和领域覆盖。

3.2.4 实验结果示例

多语言翻译结果示例如表 10 所示（以 GLM4 模型的藏汉翻译为例），其中“+XBMU 训练后”的模型输出在减少语义漂移和语序错乱方面表现最佳。

表 10 翻译结果示例

Table 10 Examples of translation results

语料类型	藏语文本	人工参考译文	模型输出文本
Base	ཞིན་ཅིན་ཚགས་པར་ལ་ “ངག་རྒྱུན་ཁྲོད་ཀྱི་བྱང་མེད་དྲ་བའི་ཉེན་རྟགས་” ཟེར་བ་ ཅིང་ཕྱིར་ཡིན། ཁྱོད་ནི་དྲ་བྱའི་ཉེན་རྟགས་མ་ཡིན་ནམ།	新京报：为什么叫“传说 中的女网警”，你不就是 网警吗？	《人民日报》上所说的“网络 空间中的女性安全风险”到底 指的是什么？难道你不是网 络安全风险吗？
+MMDS 训练 后	ཞིན་ཅིན་ཚགས་པར་ལ་ “ངག་རྒྱུན་ཁྲོད་ཀྱི་བྱང་མེད་དྲ་བའི་ཉེན་རྟགས་” ཟེར་བ་ ཅིང་ཕྱིར་ཡིན། ཁྱོད་ནི་དྲ་བྱའི་ཉེན་རྟགས་མ་ཡིན་ནམ།	新京报：为什么叫“传说 中的女网警”，你不就是 网警吗？	新疆日报中所谓的‘女性网络 监管者’是怎么回事？难道你 是网络监管者吗？
+XBMU 训练 后	ཞིན་ཅིན་ཚགས་པར་ལ་ “ངག་རྒྱུན་ཁྲོད་ཀྱི་བྱང་མེད་དྲ་བའི་ཉེན་རྟགས་” ཟེར་བ་ ཅིང་ཕྱིར་ཡིན། ཁྱོད་ནི་དྲ་བྱའི་ཉེན་རྟགས་མ་ཡིན་ནམ།	新京报：为什么叫“传说 中的女网警”，你不就是 网警吗？	新京报：为什么叫做“传说中 的女警察”？你不是网络警察 吗？

4 数据价值

本数据集在多语言自然语言处理任务中具有潜在应用价值,可为低资源语言的机器翻译提供重要训练资源,特别是在藏语、维吾尔语和蒙古语等语言对任务中,能够显著提升翻译精度和流畅性。XBMU-MC 涵盖的多领域内容,也可为跨语言信息检索、文本分类等任务提供数据支持。为确保翻译质量,本数据集在筛选过程中剔除了较多汉蒙、汉维低质量样本,致使这两种平行语料的数据量偏低,存在明显的语言对比比例失调问题。目前人工标注和校验成本较高,制约了大规模语料扩充的可行性。后续工作将重点探索基于少数民族语言基座模型的智能标注技术,通过构建专门的汉藏、汉蒙、汉维智能体实现自动化质量评估和标注流程,从而大幅降低人工成本。

5 数据使用方法和建议

使用 XBMU-MC 数据集时,应针对不同任务,合理选择微调策略与超参数。本数据集也可与其他双语开源平行语料库合并使用,利用数据集中丰富的平行语料对,引导模型学习更精确的语言对齐与翻译。若将本数据集用于多轮对话、生成式问答等更复杂的场景,可将 `instruction` 字段进行适当扩展,构建更具交互性的语言任务,适应大语言模型的复杂任务环境。数据使用须严格遵守知识产权及伦理规范,仅限于学术研究,请勿用于商业用途。

6 AIGC 使用声明

在撰写本手稿期间,作者使用了 ChatGPT-4 进行语法检查、语言润色和语句流畅度优化。所有生成内容均由作者进行修改和确认,作者对论文的整体内容、数据真实性及学术诚信负全部责任。

作者分工职责

严琦栋(2000—),男,河南南阳市人,硕士研究生,研究方向为自然语言处理。主要承担工作:数据收集、数据校对、质量审核、数据集构造、论文撰写。

马宁(1981—),男,宁夏吴忠市人,博士,教授,研究方向为自然语言处理。主要承担工作:数据集构建指导、数据质量控制、论文写作指导。

巴桑珠扎(2003—),男,西藏自治区那曲市人,本科生,专业为数据科学与大数据技术。主要承担工作:数据校对、质量审核。

白玛曲扎(2003—),男,西藏自治区日喀则市人,本科生,专业为数据科学与大数据技术。主要承担工作:数据校对、质量审核。

艾科拜·依米提(2003—),男,新疆维吾尔自治区轮台县人,本科生,专业为数据科学与大数据技术。主要承担工作:数据校对、质量审核。

麦迪那木·吾斯曼(2002—),女,新疆维吾尔自治区喀什地区莎车县人,本科生,专业为数据科学与大数据技术。主要承担工作:数据校对、质量审核。

木巴热克·阿布力克木(2001—),女,新疆阿克苏市人,本科生,专业为数据科学与大数据技术。主要承担工作:数据校对、质量审核。

苏日娜(1999—),女,内蒙古自治区赤峰市人,硕士研究生,研究方向为语言学及应用语言学。主要承担工作:数据校对、质量审核。

阿力娅（2000—），女，内蒙古自治区乌兰察布市人，硕士研究生，研究方向为语言学及应用语言学。主要承担工作：数据校对、质量审核。

参考文献

- [1] 张奇, 桂韬, 黄萱菁. 自然语言处理导论[M]. 北京: 电子工业出版社, 2023. [ZHANG Q, GUI T, HUANG X J. Introduction to natural language processing[M]. Beijing: Publishing House of Electronics Industry, 2023.]
- [2] 李洪政, 冯冲, 黄河燕. 稀缺资源语言神经网络机器翻译研究综述[J]. 自动化学报, 2021, 47(06): 1217–1231. DOI:10.16383/j.aas.c200103. [LI H Z, FENG C, HUANG H Y. A Survey on Low-resource Neural Machine Translation[J]. Acta Automatica Sinica, 2021, 47(06): 1217–1231. DOI:10.16383/j.aas.c200103.]
- [3] 朱玉, 李枫. 跨语言信息检索中的多语言处理技术研究[J]. 信息记录材料, 2024, 25(09): 98–101. DOI:10.16009/j.cnki.cn13-1295/tq.2024.09.045. [ZHU Y, LI F, et al. Research on multilingual processing technologies in cross-language information retrieval[J]. Information Recording Materials, 2024, 25(09): 98–101. DOI:10.16009/j.cnki.cn13-1295/tq.2024.09.045.]
- [4] ZHUANG W, GAO G, SUN Y. TiKG-30K: a Tibetan knowledge graph dataset based on representation learning[C]. Proceedings of the 22nd Chinese National Conference on Computational Linguistics, Harbin, China, 2023.
- [5] 翁戡, 邢天娇, 叶旭明, 等. 中-蒙-藏-维文多文档摘要数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-26). DOI:10.11922/11-6035.csd.2024.0038.zh. [WENG Y, XING T J, YE X M, et al. A dataset of Chinese-Mongolian-Tibetan-Uyghur multi-document summaries[J/OL]. China Scientific Data, 2024, 9(4). (2024-12-26). DOI:10.11922/11-6035.csd.2024.0038.zh.]
- [6] 周毛克, 俄见才让, 道吉才旦, 等. TibNER:藏文命名实体识别数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-20). DOI:10.11922/11-6035.csd.2024.0041.zh. [ZHOU M K, E J C R, DAO J C D, et al. A dataset of Tibetan named entity recognition — TibNER[J/OL]. China Scientific Data, 2024, 9(4). (2024-12-20). DOI:10.11922/11-6035.csd.2024.0041.zh.]
- [7] 黄硕, 闫晓东, 田金超. TN-SUM:藏文文本摘要数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-30). DOI:10.11922/11-6035.csd.2024.0004.zh. [HUANG S, YAN X D, TIAN J C. TN-SUM: Tibetan text summarization dataset[J/OL]. China Scientific Data, 2024, 9(4). (2024-12-30). DOI:10.11922/11-6035.csd.2024.0004.zh.]
- [8] 申影利, 包乌格德勒, 赵小兵. 蒙汉机器翻译校正数据集[J/OL]. 中国科学数据, 2022, 7(2). (2022-06-07). DOI:10.11922/11-6035.csd.2021.0102.zh. [SHEN Y L, BAO W G D L, ZHAO X B. A dataset of Mongolian-Chinese machine translation correction[J/OL]. China Scientific Data, 2022, 7(2). (2022-06-07). DOI:10.11922/11-6035.csd.2021.0102.zh.]
- [9] 冯韬, 李淼, 曹宜超, 等. 汉维可比语料数据集[J/OL]. 中国科学数据, 2020, 5(1). (2019-06-02). DOI:10.11922/csdata.2019.0010.zh. [FENG T, LI M, CAO Y C, et al. A Chinese-Uighur comparable corpus[J/OL]. China Scientific Data, 2020, 5(1). (2019-06-02). DOI:10.11922/csdata.2019.0010.zh.]
- [10] 常润, 陈波, 赵小兵. 基于语义信息扩充的汉藏短语翻译语料 SECT[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-24). DOI:10.11922/11-6035.csd.2024.0022.zh. [CHANG R, CHEN B, ZHAO X B. SECT:

- Semantic Enriched Chinese-Tibetan Phrase Translation Corpus[J/OL]. China Scientific Data, 2024, 9(4). (2024-12-24). DOI:10.11922/11-6035.csd.2024.0022.zh.]
- [11] 庞仙, 陈波, 赵小兵, 等. 基于新闻文本的汉藏新词语数据集(2017–2022)[J/OL]. 中国科学数据, 2024, 9(4). (2024-04-11). DOI:10.11922/11-6035.csd.2024.0001.zh. [PANG X, CHEN B, ZHAO X B, et al. A Dataset of Chinese-Tibetan New Words Based on News Texts (2017–2022)[J/OL]. China Scientific Data, 2024, 9(4). (2024-04-11). DOI:10.11922/11-6035.csd.2024.0001.zh.]
- [12] GOYAL N, GAO C, CHAUDHARY V, et al. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 522–538. DOI:10.1162/tac1_a_00474.
- [13] ARTETXE M, SCHWENK H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 597–610. DOI:10.1162/tac1_a_00288.
- [14] Tiedemann J. Parallel Data, Tools and Interfaces in OPUS [C]. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, 2012.
- [15] 田永红, 章钧津, 宋哲煜. 结合噪声数据增强的蒙汉伪平行语料库的构造[J]. 计算机工程与科学, 2025, 47(4): 751–760. DOI:10.3969/j.issn.1007-130X.2025.04.018. [TIAN Y H, ZHANG J J, SONG Z Y. Construction of Mongolian-Chinese pseudo-parallel corpus enhanced by noisy data[J]. Computer Engineering and Science, 2025, 47(4): 751–760. DOI:10.3969/j.issn.1007-130X.2025.04.018.]
- [16] 李林霞, 陈波, 周毛克, 等. 面向低资源语言机器翻译的平行语料句对齐评分[J]. 数据分析与知识发现, 2025, 9(1): 121–132. DOI:10.11925/infotech.2096-3467.2024.0065. [LI L X, CHEN B, ZHOU M K, et al. Parallel sentence alignment scoring for low-resource language machine translation[J]. Data Analysis and Knowledge Discovery, 2025, 9(1):121–132. DOI:10.11925/infotech.2096-3467.2024.0065.]
- [17] 魏楚元, 张鑫贤, 王致远, 等. 多注意力机制的文本摘要事实一致性评估模型[J]. 计算机工程与应用, 2023, 59(7): 163–170. DOI:10.3778/j.issn.1002-8331.2208-0167. [WEI C Y, ZHANG X X, WANG Z Y, et al. Factual Consistency Assessment Evaluation Model for Text Summarization Based on Multi-Attention Mechanism[J]. Computer Engineering and Applications, 2023, 59(7): 163–170. DOI:10.3778/j.issn.1002-8331.2208-0167.]
- [18] LOMMEL A, USZKOREIT H, BURCHARDT A. Multidimensional quality metrics (MQM): a framework for declaring and describing translation quality metrics[J]. Tradumàtica Tecnologies de la Traducció, 2014(12): 455 – 463. DOI:10.5565/rev/tradumatica.77.

论文引用格式

严琦栋, 马宁, 巴桑珠扎, 等. XBMU-MC: 多语言平行语料库[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-26). DOI: 10.11922/11-6035.csd.2025.0090.zh.

数据引用格式

严琦栋, 巴桑珠扎, 白玛曲扎, 等. XBMU-MC: 多语言平行语料库[DS/OL]. V3. Science Data Bank, 2025. (2025-07-25). DOI: 10.57760/sciencedb.25100.

XBMU-MC: A Multilingual Parallel Corpus

YAN Qidong^{1,2}, MA Ning^{1,2*}, Basangzhuzha³, Baimaquzha³, Akbar Yimit³,
Madinam Osman³, Mubarak Ablikim³, Surina⁴, Aliya⁴

1. Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou 730030, P.R. China

2. Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Gansu Province, Northwest Minzu University, Lanzhou 730030, P.R. China

3. School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou 730030, P.R. China

4. Faculty of Chinese language, Northwest Minzu University, Lanzhou 730030, P.R. China

*Email: maning@xbmu.edu.cn (MA Ning)

Abstract: Multilingual parallel corpus is a basic resource for natural language processing to break through language barriers and improve the efficiency and quality of cross-language tasks. To address the relative scarcity of parallel corpora for low-resource languages (Chinese-Tibetan, Chinese-Uyghur, Chinese-Mongolian), this study constructs a parallel corpus—XBMU-MC (Northwest Minzu University-Multilingual Corpus) for multilingual machine translation and cross-language information retrieval tasks. The original corpus consists of manually constructed specific bilingual texts and web-crawled publicly available multilingual data. The manually constructed corpus consists of Chinese texts written by researchers in the fields of culture, science and technology, and society, and the target language translations are obtained through machine translation and manual proofreading; while the web-crawled corpus captures the original texts from mainstream Tibetan, Mongolian, and Uyghur news websites to expand the scale of the corpus and the scope of the domains covered. After that, data enhancement techniques are used to expand the existing corpus, and the data are strictly screened to ensure the alignment consistency and translation accuracy between the source and target languages for each parallel corpus pair. Finally, 21,579 high-quality samples were selected and stored in JSON format, including three attributes: instruction, input and output. Finally, the multilingual translation comparison experiments between the XBMU-MC dataset and the open-source MMDS dataset based on the GLM4-9B, Qwen3-8B, and Deepseek-R1-8B models show an average increase in BLEU scores of 21.16, 20.51, and 15.37 percentage points. This dataset can help mitigate the problem of insufficient parallel corpus for low-resource languages to a certain extent, effectively support model training and multilingual translation tasks, and can also be used as a corpus basis for cross-lingual large language models in cross-language alignment, instruction fine-tuning and intelligent information processing applications in ethnic languages, helping to improve the performance of related tasks.

Keywords: Chinese-Tibetan; Chinese-Mongolian; Chinese-Uyghur; parallel corpus; data enhancement

Dataset Profile

Title	XBMU-MC: A Multilingual Parallel Corpus
-------	---

Data corresponding author	MA Ning (maning@xbmu.edu.cn)
Data authors	YAN Qidong, Basangzhuzha, Baimaquzha, Akbar Yimit, Madinam Osman, Mubarak Ablikim, Surina, Aliya, MA Ning.
Data volume	56.22 MB
Data format	*.json
Data service system	https://doi.org/10.57760/sciencedb.25100
Sources of funding	National Natural Science Foundation of China Project (62466052); Gansu Province Central Guidance Local Science and Technology Development Fund Project (25ZYJA034); Fundamental Research Funds for the Central Universities of Northwest Minzu University (31920250008).
Dataset composition	The dataset includes 21,579 json-formatted data entries, with each json object consisting of three attributes: instruction, input, and output.