# 藏文信息技术人工智能重点实验室 CCMT2022 评测报告

杨丹[1,2]，唐超超[1,2]，仁青卓玛[1,2]，拥措[1,2]*，尼玛扎西[1,2]

（1.西藏大学 信息科学技术学院，西藏 拉萨 850000; 2.西藏大学 a. 西藏自治区藏文信息技术人工智能重点实验室; b. 藏文信息技术教育部工程研究中心，西藏 拉萨 850000）

**摘要：**本文详细介绍了西藏自治区藏文信息技术人工智能重点实验室（西藏大学）参加第十八届全国机器翻译大会（CCMT 2022）（China Conference on Machine Translation, CCMT）的藏汉机器翻译评测情况。本次测评采用了谷歌 Transformer 神经网络机器翻译架构作为基线翻译模型，主要利用数据增强的方式对藏汉平行语料进行扩充、优化藏汉神经机器翻译所用到的词表并探索跨语言预训练模型中的联合词表对翻译性能的影响，最终使用了一种融合跨语言预训练模型 mRASP 与改进后的绿色联合词表的方法提交了藏汉神经机器翻译的主系统。对比系统分别是基于 ALBERT 预训练语言模型和基于 transformer-big 的翻译系统。与传统的基于预训练语言模型的藏汉神经机器翻译相比，该方法可以有效提高藏汉机器翻译性能。

**关键词：**跨语言预训练模型；藏汉神经机器翻译；mRASP；词表；ALBERT

**中图分类号：**TP 391　　**文献标志码：** A

## 1 引言

本文详细介绍了西藏自治区藏文信息技术人工智能重点实验室（西藏大学）参加第十八届全国机器翻译大会（CCMT 2022）（China Conference on Machine Translation, CCMT）的藏汉机器翻译评测情况。本次测评采用了谷歌 Transformer[1]神经网络机器翻译架构作为基线翻译模型，在数据预处理方面，针对评测方提供的平行语料主要采用全角半角转换，非法字符删除，长度比限制等等。为了提高模型的效果，进行了语料的扩充。利用评测提供的平行语料和汉语单语语料，通过同义词替换和回译的数据增强方式生成伪平行语料。同时考虑到伪平行语料的质量问题，对伪平行语料进行了过滤。在模型训练方面，用 Transformer-big 模型训练 NMT 模型，并训练了融合跨语言预训练模型 mRASP[2]与改进后的绿色联合词表的藏汉神经机器翻译，根据验证集的表现选择最优模型。在译文输出过程中，对长度惩罚因子参数进行了调优。

**基金项目：**科技部重点研发计划重点专项(2017YFB1402200)；西藏自治区科技创新基地自主研究项目(XZ2021JR002G)；西藏大学珠峰学科建设计划项目（zf22002001）

* **通信作者：**yc@utibet.edu.cn

在本次 CCMT 2022 双语翻译评测任务中，西藏自治区藏文信息技术人工智能重点实验室提交了藏汉机器翻译的一个评测主系统和两个对比系统的翻译结果。

# 2 数据处理

## 2.1 数据预处理

（1）由于评测方提供的平行语料的文件格式不一致，所以首先对平行语料进行提取，然后进行符号标准化。具体包括全角半角转化，删除非法字符，大小写转换以及中文化繁为简等。

（2）分词处理。汉语使用北大 pkuseg[3]分词，藏语使用 TIP-LAS[4]分词。数据处理后，为了缩减词表和解决集外词 OOV（Out of Vocabulary）问题，采用 subword-nmt[5]训练 BPE 并应用于语料，分别生成藏语词表和汉语词表，藏语和汉语子词规模均为 32k。随后对词表进行优化，并与优化前进行对比。

（3）对所有数据进行长度比过滤。过滤藏汉双语句对中长度过长或过短的句对。经过清洗，最终得到 115 万条数据，然后通过随机抽样方法从中随机抽取 2000 条数据划分为验证集和测试集各 1000 条。

（4）为了提高模型泛化能力，把训练集中与验证集和测试集重复的 106 条句对删除。

## 2.2 数据增强

数据增强是提升藏汉机器翻译的有效途径。因此为了提高机器翻译模型的性能，首先进行数据增强。本文主要通过同义词替换和回译的方式扩充语料。

（1）同义词替换

在训练集中随机抽取 15 万条数据采用同义词替换的数据增强方式扩充语料。在进行藏语的同义词替换时，使用 50 万条藏语单语语料训练 word2vec 模型。其中 word2vec 包含两种训练算法，此处选用 skip-gram 方法。从句子中根据替换率分别为 0.08、0.15 的频率随机选择非停用词进行替换。汉语语料借助中文近义词工具包 Synonyms，从句子中根据替换率分别为 0.08、0.15 的频率随机选择非停用词进行替换。其中 word2vec 模型的词向量可视化如图 1 所示。

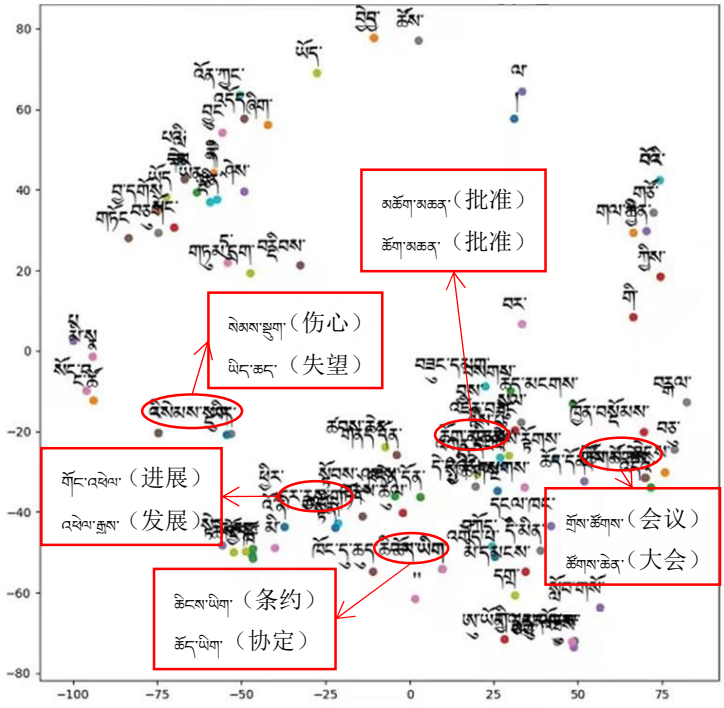


图 1 词向量可视化

Fig.1 Word vector visualization

从图 1 可以看出，意思相近的词被映射到了向量空间中相近的位置。

（2）回译

回译的主要思想是：利用目标语言-源语言翻译模型（反向翻译模型）来生成伪双语句对，用于训练源语言-目标语言翻译模型（正向翻译模型）。首先，使用双语数据训练汉藏翻译模型，然后通过该模型将汉语句子翻译为藏语，从而得到大量的藏语-汉语伪双语句对，此处用到的汉语单语语料为 20 万条。然后，将回译得到的伪双语句对和真实双语句对混合，训练得到最终的藏汉翻译模型。

虽然回译获得的藏语语料的质量不是很好，但是其对应的汉语语料，即目标语言端仍然是真实数据，可以使解码器训练得更加充分，也可以提高模型的健壮性。

经过上述方法进行数据增强后，再进行语料过滤。最终的语料规模如表 1 所示。

表 1 语料规模

Tab.1 Size of corpus

| 数据集 | 数据增强方式 | 句对数 |
|---|---|---|
| 训练集 | - | 1153421 |
| | 同义词替换 | 1373525（+220104） |
| | 回译 | 1558917（+185392） |
| 验证集 | - | 1000 |
| 测试集 | - | 1000 |

# 3 方法介绍

## 3.1 VOLT 词表

VOLT[6]通过考虑语料库熵和词汇量大小在多项式时间内给出合适的词汇量。具体而言：首先借用了经济学中边际效用的概念，使用 MUV（词汇的边际效用）作为评估方法。形式上，MUV 被定义为熵对词汇量大小的负导数；然后将目标转向在可处理的时间复杂度中最大化 MUV，将离散优化目标重新表述为最优转移问题，可以通过线性规划在多项式时间内求解；最后从最佳转移矩阵中生成词表。

使用 VOLT 优化词表可以找到具有更高的 BLEU 值、更小尺寸的性能良好的词汇表，并且在模型的训练过程中，可以缩短其训练所需的时间。因此本文根据藏文本身特点，采用 BPE 后运用 VOLT 构建词表，并与 BPE 方法进行对比。

## 3.2 基于 mRASP 的藏汉神经机器翻译方法

利用大量比较容易获得的数据来预训练模型，然后在具体应用场景利用少量数据微调，已经成为自然语言处理的新成功范式。例如 BERT[7]利用大规模数据进行预训练后，在自然语言理解的 11 项任务上少量微调就能取得很好的成绩。在机器翻译任务上，mRASP 设计了一个通用的预训练模型。为了区分不同的语言对，会在句子前添加人工语言标记来表示源语言和目标语言。神经网络结构采用 Transformer，随机对齐替换技术（Random Aligned Substitution, RAS）的引入使得不同语言的同义词之间共享相同的上下文，从而进一步拉近了不同语言之间同义词的表示。

由于跨语言预训练语言模型 mRASP 包含的语义知识较多，所以本文将使用一种融合跨语言预训练模型 mRASP 与改进后的联合词表的方法训练藏汉神经机器翻译，从而进一步提高藏汉神经机器翻译的质量。

# 4 实验

## 4.1 实验环境

在本次机器翻译评测中，本团队使用的操作系统为 Ubuntu，深度学习框架为 pytorch 1.10.1，机器翻译框架为 fairseq 0.10.2。

## 4.2 实验参数

本次测评采用 Fairseq 系统的 transformer-big 模型，使用 Adam 梯度优化算法来训练得到最终的模型参数，其中$\beta_1$=0.90，$\beta_2$=0.98。为了防止过拟合，将 dropout 参数设置为 0.1，clip norm 参数设置

为 10，weight decay 参数设置为 0.01。其余实验参数设置如表 2 所示。

表 2 实验参数设置

Tab.2 Experimental parameter settings

| 参数名称 | 参数选择 |
|---|---|
| encoder_layers | 6 |
| decoder_layers | 6 |
| attention_heads | 16 |
| embed_dim | 1024 |
| ffn_embed_dim | 4096 |
| optimizer | adam |
| dropout | 0.1 |
| label_smoothing | 0.1 |
| warmup updates | 4000 |
| lr | 0.0003 |

## 4.3 基线实验的选择

首先用 BPE、以及 VOLT 构建词表的方法来构建藏汉神经机器翻译模型，然后对比两个模型的性能，实验结果如表 3、表 4 所示。

表 3 词表构建规模对比

Tab.3 Comparison of vocabulary construction size

| 词表构建 | 藏文 | 中文 |
|---|---|---|
| BPE | 33325 | 42911 |
| VOLT | **10360** | **22518** |

表 4 验证集实验效果对比

Tab.4 Experimental effect comparison of valid set

| 词表构建 | BLEU |
|---|---|
| BPE | 49.89 |
| VOLT | **52.26** |

实验表明，由 VOLT 构建词表可以在一定程度上缩小词表的规模，而词表大小也会影响下游任务表现，在使用 VOLT 构建词表后，藏汉翻译提高了 2.37 个 BLEU 值。所以本文选择 VOLT 构建词表为基准模型。

## 4.4 实验结果及分析

（1）长度惩罚因子

为了适应验证集短句较多的情况，在藏汉神经机器翻译上分析了长度惩罚因子α对 BLEU 值的影响。解码时，束搜索大小设置为 5，采用不同值的长度惩罚因子进行实验。实验结果如表 5 所示。

表 5 长度惩罚因子对藏汉翻译 BLEU 值的影响

Tab.5 Effect of length penalty factor on the BLEU values of Tibetan-Chinese translation

| α | 0.1 | 0.2 | 0.3 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|---|
| BLEU | 52.19 | 52.25 | **52.26** | 52.15 | 51.89 | 51.74 | 51.67 |

由表 5 可知，藏汉翻译适合的α为 0.3，合适的长度惩罚因子α会对 BLEU 值产生正面影响，过大或者过小的长度惩罚因子α都会影响翻译性能。

（2）mRASP 跨语言预训练模型

由于 mRASP 获取的 32 种语言的训练集大小不均，所以先把所有语料混合起来，通过过采样（Over-sampling）去平衡词汇量，保持词汇表中词汇的最低频度为 20。然后通过 BPE 切分得到包含 32 种语言的联合词表，其规模为 64808。为了扩大藏语和汉语的词表占比，需要把藏语词表和汉语词表合并到原有的词表中。本文采取 4 种方法合成词表：

①将 BPE 切分产生的藏语、汉语词表分别合并到原有词表中；

②将藏语、汉语语料混合起来，通过采样平衡词汇量，然后 BPE 切分得到藏语、汉语联合词表，将其合并到原有词表中；

③将 VOLT 切分产生的藏语、汉语词表分别合并到原有词表中；

④将藏语、汉语语料混合起来，通过采样平衡词汇量，然后 VOLT 切分得到藏语、汉语联合词表，将其合并到原有词表中；

mRASP 提供了两个 32 个语言对的模型，其中 w/o model 模型不包括对齐信息，w/ model 模型包括 RAS 对齐信息。由于 ALBERT[8]是轻量级 BERT，通过权值共享和矩阵分解减少参数，降低了空间复杂度，且使用遮蔽语言模型 MLM 和 Transformer 的编码器来生成深度的双向语言特征向量，所以此处将其作为对比实验。本文在训练好藏语 ALBERT 模型后，先根据下游任务调整 ALBERT 模型参数，然后将微调好的模型参数迁移到 Transformer 的编码器端。实验对比如表 6、表 7 所示。

表 6 mRASP 词表构建规模对比

Tab.6 Comparison of the scale of mRASP vocabulary construction

| 词表 | 藏文 | 中文 |
|---|---|---|
| 基线(transformer+VOLT) | 10360 | 22518 |
| ALBERT+基线 | 10360 | 22518 |
| mRASP+BPE | 128142 | 128142 |
| mRASP+基线 | **85330** | **85330** |
| mRASP（joint_b） | 108243 | 108243 |
| mRASP（joint_v） | 92768 | 92768 |

表 7 mRASP 实验结果对比

Tab.7 Comparison of the results of mRASP experiments

| 方法 | 藏→汉 | |
|---|---|---|
| | （w/o model） | （w/model） |
| 基线(transformer+VOLT) | 52.26 | |
| ALBERT+基线 | 52.78 | |
| mRASP+BPE | 52.98 | 53.67 |
| mRASP+基线 | 54.71 | **55.69** |
| mRASP（joint_b） | 53.34 | 54.09 |
| mRASP（joint_v） | 53.56 | 54.25 |

由表 6、表 7 可知，mRASP+基线的方法在训练集上获得了最好的效果。相比基线方法，其他方法的翻译性能明显提升。对比使用 ALBERT+基线模型，mRASP 不论采用何种方式构建词表，其模型的性能均得到了不同程度的提升，可以看到引入跨语言预训练模型可以有效地改善模型的翻译效果。在 mRASP 上融合 BPE 联合词表相较于 mRASP+BPE 来说，在一定程度上缩小了词表规模，在藏汉翻译上提高了 0.42 个 BLEU 值；但 mRASP 上融合 VOLT 联合词表相较于 mRASP+基线模型来说，反而增大了词表规模，降低了翻译性能。这可能是由于 VOLT 优化词表所依靠的是信息熵，而藏语和汉语的信息熵差别较大而导致的。

由表 6、表 7 的词表规模和实验结果绘制的折线图如图 2 所示。

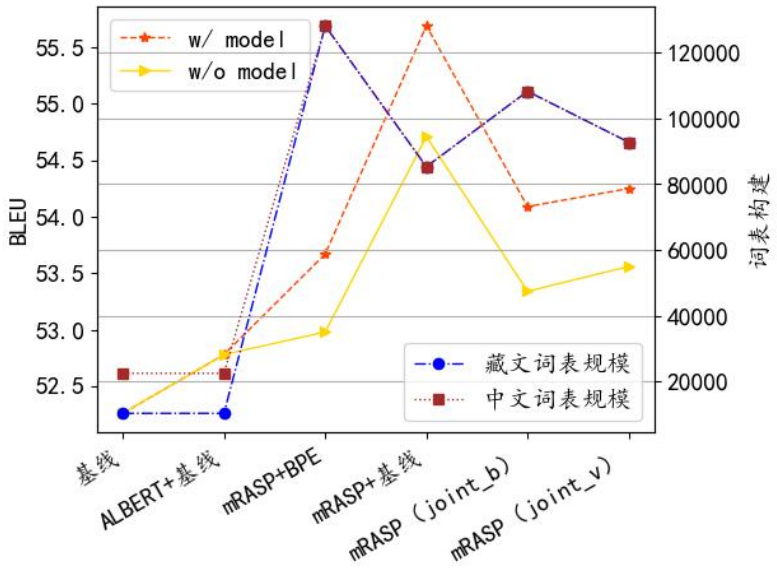


图 2 藏汉翻译对比

Fig.2 Tibetan-Chinese Translation Comparison

图 2 直观地展示了在 mRASP 模型上，mRASP+基线模型的实验效果是最优的，达到了最佳的翻译效果，且在 w/ model 模型都优于 w/o model 模型。

## 4.5 译文分析

为了直观地对比藏汉机器翻译的改进效果，从 1000 条测试集中随机抽取 1 条语句绘制成表。藏汉翻译效果对比如表 8 所示。

表 8 藏汉翻译结果对比

Tab.8 Comparison of Tibetan-Chinese translation results

| | |
|---|---|
| 实例 | "གཡོ་འགུལ་སྐྱ་ཙམ་ཡང་མེ་གཏོང་བ་གཉིས་"མཐའ་འཁོར་དང་། དབང་ཆ་འདྲ་མཉམ་དང་། གོ་སྐབས་འདྲ་མཉམ། སྒྲིག་སྡོལ་འདྲ་མཉམ་བཅས་མཐའ་འཁོར། |
| 基线 | 坚持“毫不动摇“两个毫不动摇”，坚持权利平等、机会均等、规则平等； |
| ALBERT+基线 | 坚持“毫不动摇”，坚持权平等、机会均等、规则平等； |
| mRASP+基线 | 坚持“两个毫不动摇”，坚持权利平等、机会平等、规则平等； |
| 参考译文 | 坚持“两个毫不动摇”，坚持权利平等、机会平等、规则平等； |

总体来看，3 个译文都与参考译文意思一致，但是基线将“毫不动摇”多译，ALBERT+基线模型将“གཡོ་འགུལ་སྐྱ་ཙམ་ཡང་མེ་གཏོང་བ་གཉིས་”译为“毫不动摇”，漏译“两个”。“དབང་ཆ་འདྲ་མཉམ་”漏译为“权平等”。mRASP+基线中准确无误地把“གཡོ་འགུལ་སྐྱ་ཙམ་ཡང་མེ་གཏོང་བ་གཉིས་”翻译成“两个毫不动摇”，明显更符合原意。

# 5 总结

本文介绍了 CCMT 2022 的语言测评项目的藏汉方向上使用的主要方法和技术。评测中我们采用

Transformer 神经机器翻译模型作为翻译系统的主要技术，结合使用 VOLT 改进词表、探索联合词表对翻译性能的影响，并在 mRASP 跨语言预训练模型上进行融合来提升翻译质量，解码时使用合适的长度惩罚因子。实验表明，这些方法能够明显提高藏汉翻译的质量。

由于时间有限，本次测评中还有很多工作尚未进行完善、一些方法尚未尝试，在今后的研究学习中，希望对机器翻译的相关技术有更深一步的认识。

## 参考文献


[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30:6000-6010.

[2] Lin Z, Pan X, Wang M, et al. Pre-training multilingual neural machine translation by leveraging alignment information[C]//Proc of EMNLP. 2020: 2649-2663.

[3] Luo R, Xu J, Zhang Y, et al. Pkuseg: A toolkit for multi-domain chinese word segmentation[J]. arXiv preprint arXiv:1906.11455, 2019.

[4] 李亚超, 江静, 加羊吉, 等. TIP-LAS: 一个开源的藏文分词词性标注系统[J]. 中文信息学报, 2015, 29(6): 203-207.

[5] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]// Proc of ACL (Volume 1: Long Papers). 2016: 1715-1725.

[6] Xu J, Zhou H, Gan C, et al. Vocabulary learning via optimal transport for neural machine translation[C]//Proc of ACL (Volume 1: Long Papers). 2021: 7361-7373.

[7] 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.

[8] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

# Tibetan Information Technology Artificial Intelligence Key Laboratory CCMT2022 Review Report

YANG Dan[1,2]，TANG Chaochao[1,2]，RENQING Zhuoma[1,2]，YONG Cuo[1,2]*，

NIMA Zhaxi[1,2]

(1. School of Information Science and Technology, Tibet University, Lhasa, Tibet,850000, China;2. a. State Key Laboratory of Artificial Intelligence for Tibetan Information Technology in Tibet Autonomous Region, b. Ministry of Education Engineering Research Center for Tibetan Information Technology, Tibet University, 850000, China)

**Abstract:** This paper details the evaluation of Tibetan-Chinese machine translation at the 18th China Conference on Machine Translation (CCMT 2022) by the Key Laboratory of Artificial Intelligence for Tibetan Information Technology in Tibet Autonomous Region (Tibet University). The evaluation used the Google Transformer neural network machine translation architecture as the baseline translation model, mainly used the data enhancement method to expand the Tibetan-Chinese parallel corpus, optimized the vocabulary used in Tibetan-Chinese neural machine translation and explored the impact of the joint vocabulary in the cross-lingual pre-training model on the translation performance, and finally, a method combining cross-language pre-training model mRASP and improved green joint vocabulary is used to submit the main system of Tibetan-Chinese neural machine translation. The comparison system is based on ALBERT pre-training language model and transformer-big translation system. Compared with the traditional Tibetan-Chinese neural machine translation based on pre-training language model, this method can effectively improve the performance of Tibetan-Chinese machine translation.

**Keywords:** Cross-language pre-training model; Tibetan-Chinese neural machine translation; mRASP; vocabulary; ALBERT