CCMT2022 低资源藏汉机器翻译评测报告

严松思[1,2]，汪超[1,2&]，珠杰[1,2*]

（1. 西藏大学信息科学技术学院，西藏 拉萨 540000；
2. 省部共建西藏信息化协同创新中心，西藏 拉萨 540000）

摘要：本文介绍了实验室参加第十八届全国机器翻译大会（CCMT2022）机器翻译评测中的藏汉日常用语机器翻译评测项目的情况。本报告主要介绍参赛的藏汉神经网络机器翻译系统以及该系统所采用的方法及该系统在开发集和测试集上的性能。

关键词：机器翻译；模型集成；双向训练策略；mBART 模型

中图分类号：TP391　　**文献标志码：**A

1. 引言

本文介绍了本实验室所参加第十八届全国机器翻译大会（CCMT 2022）（China Conference on Machine Translation，简称 CCMT）的藏汉机器翻译技术评测的主要情况。为了提高模型的效果，得到质量更好的生成数据，本次评测采用 Transformer、mBART 模型神经网络机器翻译架构，使用模型集成策略以及双向训练策略，同时给出不同设置下评测系统在评测数据集上的性能表现，并进行了对比和分析。

2. 数据处理

2.1 数据预处理

本次评测语料来源是 CCMT2022 提供的藏汉政府文献机器翻译语料，由于原始的平行语料未做任何处理，这对翻译质量有着很大的影响。因此需对所提供的原始语料进行预处理，处理过程如下：（1）全角字符转换为半角字符；（2）对藏语和汉语语料分别进行分词处理，其中藏语使用中科院提供的分词软件，汉语使用 Jieba 分词；（3）对数据进行子词长度及长度比过滤：最大子词长度比设置为 1，句子长度过滤区间设置为[1, 100]。评测实验中对涉及到的所有语料，包括训练集、开发集、测试集都做了相同的预处理工作。

2.2 子词化处理

BPE(字节对)[1]编码是一种简单的数据压缩形式，其分词粒度处于单词级别和字符级别之间，其将词本身的意思和词的形态变化部分分开，有效的减少了词表的数量。同时也是目前解决机器翻译任务中未登录词问题的一种普遍方法。本次评测使用 subword-nmt 工具处理藏汉翻译数据集，词表大小为 50000，并联合源语言和目标语言进行 BPE 切分处理以创建藏汉联合词表。

2.3 评测测试数据集处理

通过对本次 CCMT2022 所提供的藏文测试集进行分析，测试集中包含 10000 条数据，其大部分为长文本段落，这对翻译效果有着不利的因素，因此对每个段落进行切割是十分有必要的。

2.3.1 藏文句子边界判定

藏文是一门古老的语言，至今已有 1400 多年，藏文有着自身独特、完整的的语法体系。但对于藏文而言，句子与句子之间没有特定的分隔符，在藏语独特的标点符号中，涉及标识句子结束的标

基金项目：国家自然基金项目(62066042)；教育部人文社会科学研究项目(21YJCZH059)；2021 年西藏自治区高校人文社会科学研究项目(SK2021-24)；西藏大学提升计划项目(ZDTSJH21-07)；西藏大学培育计划项(ZDCZJH21-10)；西藏大学珠峰学科建设计划项目(zf22002001)

&同等贡献

***通信作者：**rocky_tibet@qq.com

点符号主要是楔形符，楔形符包括单垂符“།”、双楔形符“།།”和四楔形符“།།།།”。

单垂符“།”用来表示藏文的词、句子的停顿或结束，在藏文句子中主要用于句末，也可用于词或者短语之后，在功能上相当于汉语中顿号，逗号，句号，问号等标点符号。双楔形符“།།”与单垂符用法基本一致，一般比较强调句子结束。四楔形符“།།།།”用于标记大文章或一章内容结束。

藏文句子结束的地方一般有楔形符，但有楔形符的地方并不一定是藏文句子的边界，因此根据楔形词来划分藏文长文本中句子边界是不现实的。不过如今现代藏语的文章并没有完全按照上述的规则进行切分，越来越多的单垂符“།”代替了双楔形符“།།”和四楔形符“།།།།”。

2.3.2 词性规则句子边界判定法

相比较中文、英文等文字可以通过标点符号识别句子边界，而藏语主要的楔形符却因自己独特的作用难以识别句子的边界，但可通过藏文文法理论来作为依据识别句子边界。本文利用马伟珍[2]等人提出的词性规则法判定句子边界，其基本思想是：首先，读入已分词并词性标注完成的文本；其次，对数据从头到尾进行扫描，每次遇到单垂符或双楔形符，判断其前一个或后一个任意词的词性是否为连词、前一个词的词尾是否为ས或ན，前一个词的词性是否为名词、数词或状态词，或后一个词的词性是否为符号（包括藏文符号和非藏文符号）。若是，则继续扫描，否则，进行分句切割。

3. 模型介绍

3.1 Transformer 网络模型

Transformer [3]模型采用自注意力计算机制，由编码器和解码器两部分组成。其中编码器有 6 层，解码器也有 6 层，如图 1 所示。编码器是由 N=6 个相同的层组成的堆栈，每个子层的输出是 LayerNorm(x+Sublayer(x))，其中 Sublayer(x)是子层自己实现的函数。解码器结构和编码器大体一致，不过在输入序列提取特征时采用的是 Masked 多头注意力层，整个模型使用残差连接与对各层输出使用规范化来更好的优化深度网络。

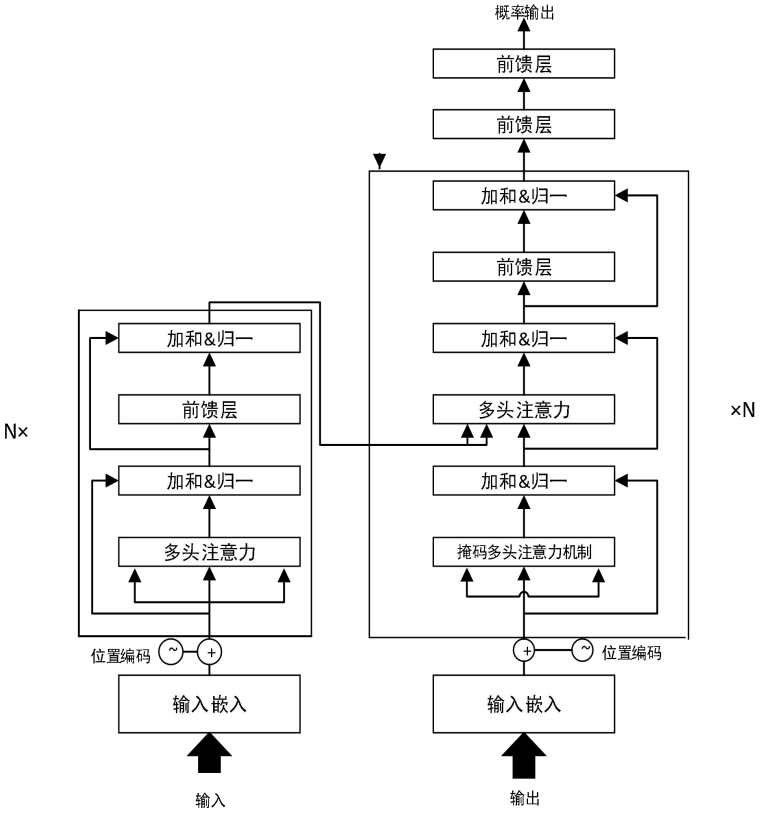

图 1 Transformer 结构图

Fig. 1 Transformer structure diagram

3.2 mBART 模型

mBART 模型的基础是 BART 模型。BART[4]是提出的一种新的预训练范式，包括两个阶段：首先原文本使用某种 noise function 进行破坏，然后使用序列到序列模型还原原始的输入文本。

BART 模型在 BERT 模型和 GPT 模型[5][6]的基础上进行改进。在 BERT 模型中，如图 2 所示，随机 token 被替换为掩码，并且文档被双向编码。由于其缺失的 token 是独立预测的，因此 BERT 不能轻易地用于生成。

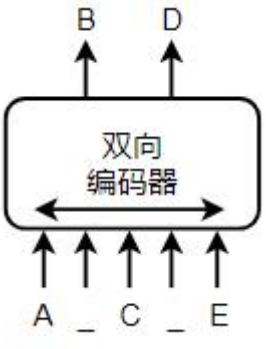

图 2 BERT 的训练方式

Fig. 2 BERT's training methods

而在 GPT 模型中，如图 3 所示，其 token 是自动回归预测的，这意味着 GPT 可以用于生成。然而，单词只能适应左向的上下文，所以它不能学习双向交互。

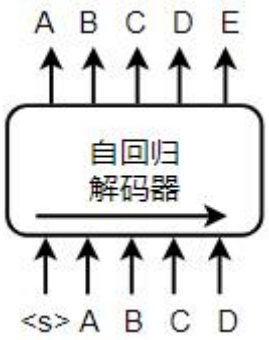

图 3 GPT 的训练方式

Fig. 3 GPT's training methods

BART 模型，如图 4 所示，结合了上述两种模型的优点，它对编码器的输入不需要与解码器的输出对齐，允许任意噪声转换。

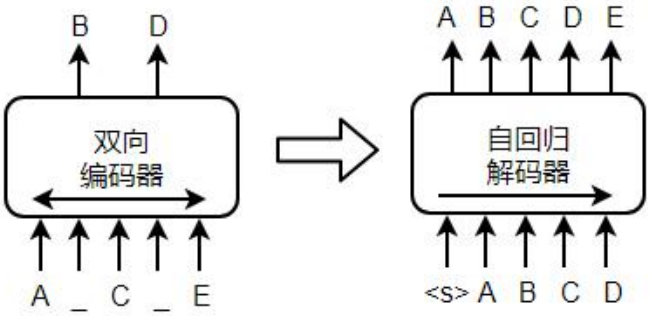

图 4 BART 的训练方式

Fig. 4 BART's training methods

mBART[7]在 BART 模型的基础上，遵循 BART 序列到序列的预训练方案。mBART-base 模型使用标准的 Transformer 架构，包括 6 层编码器和 6 层解码器。同时在编码器和解码器的基础上包括了一个额外的层归一化层。

该模型的噪声函数：在 g 中使用了两种类型的噪声，删除文本的跨度，并用 mask token 代替。其一，按照泊松分布（λ=3.5）随机抽取 token，然后对每个实例中 35%的词进行 mask。其二，对一个原始输入的不同句子进行调换顺序。

3.3 模型集成

模型集成是使用多个模型同时解码，即在解码过程中，在经过 Softmax 得到归一化的目标语言词表上的概率分布后，整合不同模型得到的概率分布，进而预测下一个词，融合各个模型的学习能力，从而提升模型的鲁棒性，增强模型的泛化能力。本文采用同一模型不同训练轮数的模型集成方法。该方法即将最后几轮训练模型结果做集成，一方面可降低随机误差，另一方面也避免了训练轮数过多带来的过拟合风险。

3.4 双向训练策略

Liang 等人[8]研究人类学习行为发现，双向语言学习可以更好的帮助神经机器翻译模型学习，在翻译任务中，“源端到目标端”和“目标端到源端”的平行数据对于任意一个方向的训练都应该有帮助，因此提出一种简单的数据控制手段来实现双向训练。

给定一个源句子 x,NMT 模型会根据之前生成的目标端单词 y<x 逐个生成每个目标单词y_t，因此，生成 y 的概率计算如下：

$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{T} p(\mathbf{y}_t \mid \mathbf{x}, \mathbf{y} < t; \theta) \tag{1}$$

其中 T 为目标序列的长度，θ为目标函数，最大似然估计的参数为$\mathcal{L}(\theta)$。

$$\mathcal{L}(\theta) = \arg\ \max_{\theta} \log\ p(\mathbf{y} \mid \mathbf{x}; \theta) \tag{2}$$

传统 Transformer 机器翻译训练的数据被定义为：

$$\vec{\mathrm{B}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N} \tag{3}$$

其中 N 是样本总数，x 和 y 分别代表源语言和目标语言句子。

双向训练技术通过预训练的形式迁移知识，并且提升下游任务的泛化性，基于双向的训练的 Transformer 机器翻译训练的数据被定义为：

$$\overleftrightarrow{\mathrm{B}} = \{(\mathbf{x}_i, \mathbf{y}_i) \cup (\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^{N} \tag{4}$$

此时，Transformer 模型的参数θ由两个方向的数据同时进行更新：

$$\overleftrightarrow{\mathcal{L}}(\theta) = \overbrace{\arg\ \max \log\ p(\mathbf{y} \mid \mathbf{x}; \theta)}^{\text{Forward: } \overrightarrow{\mathcal{L}_\theta}} + \underbrace{\arg\ \max}_{\theta} \log\ p(\mathbf{x} \mid \mathbf{y}; \theta) \tag{5}$$

在数据层面上，双向训练技术通过以下 2 个步骤实现双向更新，如图 5 所示。

（1）交换平行语料的源语料和目标语料；

（2）将交换后的训练数据集加入原本的训练数据集中。此时，训练数据集的总量会翻倍，最后将翻倍的数据集进行训练。同时，为了保证实验的公平性，实验并没有增加总的训练步数，而是选用原本总步数的 1/3 进行预训练，然后在正常翻译方向的训练数据上进行剩下 2/3 步数的更新。

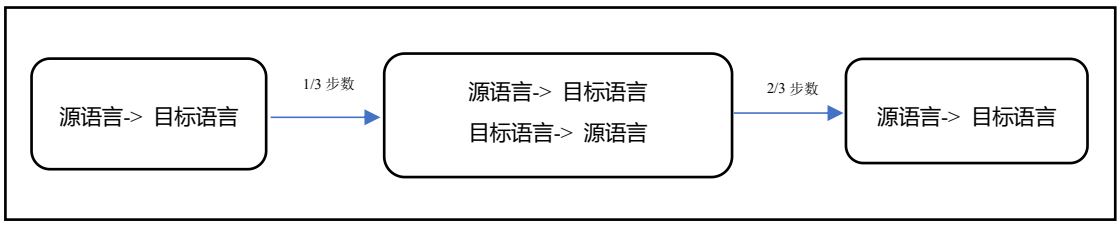

图 5 双向训练策略数据集变化

Fig. 5 bidirectional training strategy data set changes

4. 实验

4.1 实验设置

本文实验的硬件环境：操作系统为 Ubuntu20.04，GPU 为 RTX4000。使用 fairseq 开源框架进行藏汉神经机器翻译模型的搭建，本文采用 Transformer、mBART 模型,模型参数中均设置句子长度最大为 50 词，词向量维度为 512，训练样本 batch_size 为 128，网络丢弃率 Droput 为 0.1，过滤器大小设置为 2048，神经网络层数设置为 6 层，使用 Adam 优化算法，学习率分别设置为 1e-4 及 5e-4，使用 Vaswani 等描述的学习率衰减策略。解码阶段采用集束搜索策略并设置 beam width 为 5。

实验具体数据集如下所示：

表 1 数据集大小

Tab.1 Dataset size

名称	规模
训练集	1131124 句对
开发集	1000 句对
测试集	1049 句对

4.2 实验结果与分析

本文设置了 5 组实验，并通过 BLEU 值进行评价。第一组是将 Transformer 模型作为本文的基线实验，使用 CCMT 所提供的数据集进行实验。第二组是在基准实验中引用双向训练技术，通过该技术改变训练方向，扩充一倍数量的训练集，并在不改变总训练步数的前提下，首先在训练集中训练 1/3 步数，然后在正常训练方向继续训练余下的 2/3 步数（Transformer+BIT）。第三组实验是将 mBART 模型作为实验模型，保持与基线实验参数相同。第四组实验依然将 mBART 模型作为实验模型，并在实验中引用模型集成技术。第四组实验依然将 mBART 模型作为实验模型，在实验中引用双向训练技术。(mBART+BIT)

表 2 实验结果

Tab.2 Experimental results

模型	1e-4	5e-4
Transformer	27.75	33.52
Transformer+BIT	28.31	34.24
mBART	29.85	35.26
mBART+双模型集成	30.23	36.12
mBART+BIT	30.87	36.93

分析实验结果可知，在评测任务中，在采用模型平均和模型集成策略后，翻译质量较基线模型 BLEU 值提升。因此可以得出结论：

（1）使用 mBART 模型架构可以提升翻译的质量；

（2）模型集成对翻译质量提升有一定的帮助，加入两个模型集成较基线模型 BLEU[10]值提升了 0.86，0.7，这说明在解码阶段，通过整合不同模型得到的概率分布来预测下一个词，是对翻译质量有所提升的。

（3）由第二组与第五组实验结果可知,在引入双语训练策略后，第二组实验 BLEU 值相较于第一组分别提升了 0.56，0.72，第五组实验相较于第三组 BLEU 值分别提升了 0.64，0.81。这是因为双向训练策略是一个句子级且更换概率为 0.5 的 code-switch 方法，以藏文-中文中的句子{“དར་འཕྲེན་མོན་ནས་ཚ་ཤུར་རན་རེས།”->“蚕进入母龄后”}为例，在预训练阶段，重构的预训练数据中同时包含正向的{“དར་འཕྲེན་མོན་ནས་ཚ་ཤུར་རན་རེས།”->“蚕进入母龄后”}与反向的{“蚕进入母龄后”->“དར་འཕྲེན་མོན་ནས་ཚ་ཤུར་རན་རེས།”}。此时，反向的句对可以看作概率为 0.5 的句子级 switch。同时，双向训练策略可以更好的鼓励自注意力机制学习更好的双语关系，可以得到更好的双语注意力矩阵，提升双语对齐质量。

4.3 译文展示

表 3 译文展示

Tab.3 Translation Example

源语言句子	གནས་བབ་ གསར་ཏ་ དང་ ལས་འགན་ གསར་ཏ་ ར་ དེམེགས་ ནས་ ང་ཚོ་ ཡི་ ཏང་ གིས་ དུས་རྒྱུན་ དུ་ ཏང་ཡོངས་ ཀྱི་ ཟོ་མཐུན་ ཚོ་ ར་ སྐོབ་སྦྱོང་ ལ་ ཤུགས་རྒྱག་ དགོས་ ཏ་ ཡི་ འདོད་སྐུལ་ བཏང་ ཡོད །
参考译文	面对新形势新任务，我们党经常号召全党同志加强学习。
Transformer	对 新 时期 的 工作 ，我们 党 一直 呼吁 全党同志 加强 学习 。
Transformer+BIT	面对 新的 时期 的 工作 ，我们 党 一直 呼吁 全党同志 加强 学习 。
mBART	面对 新 形势 新 任务 ， 我们 党 一直 呼吁 全党同志 加强 学习 。
mBART+双模型集成	面对 新 形势 新 任务 ， 我们 党 一直 号召 全党同志 加强 学习 。
mBART+BIT	面对 新 形势 新 任务 ， 我们 党 常常 号召 全党同志 加强 学习 。

参考文献：

[1] SENNRICH R, HADDOW B, BIRCH A. 2016. Neural Machine Translation of Rare Words with Subword Units [C]// Proc of the 54th ACL. Stroudsburg, PA: ACL, 2016: 1715-1725.

[2] 马伟珍,完么扎西,尼玛扎西. 藏语句子边界识别方法[J]. 西藏大学学报(自然科学版),2012,27(02):70-76.DOI:10.16249/j.cnki.54-1034/c.2012.02.010.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In Advances in Neural Information 7 Processing Systems, pages 5998-6008.

[4] Lewis, Mike, et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." ACL 2020: 7871-7880

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. 2018. Bert: Pre-training of deep bidirectional Transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Guillaume Lample and Alexis Conneau. 2019. Crosslingual language model pretraining. arXiv preprint arXiv:1901.07291.

[7] Liu, Yinhan, et al. "Multilingual denoising pre-training for neural machine translation." TACL.2020

[8] Liang Ding, Di Wu, Dacheng Tao. Improving Neural Machine Translation by Bidirectional Training.[C] In emnlp . arXiv:2107.11572

[9] Alexis Conneau, German Kruszewski, Guillaume Lam ple, et al. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In ACL.

[10] Papineni, K., Roukos, S., Ward, T., et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318 (2002).

CCMT2022 Low Resource Tibetan-Chinese Machine

Translation Evaluation Report

YAN Songsi[1,2], WANG Chao[1,2&],ZHU Jie[1,2*]

(1. School of Information Science and Technology, Tibet University, Lhasa 540000, China; 2. Provincial and Ministerial Collaborative Innovation Centre for Informatization in Tibet, Lhasa 540000, China)

Abstract: This paper presents the participation of our lab in the machine translation evaluation project of the 18th National Conference on Machine Translation (CCMT2022) for the evaluation of Tibetan-Chinese daily expression. It focuses on the participating Tibetan-Chinese neural network machine translation system as well as the methodology used in the system and the performance of the system on the development and test sets.

Keywords: Machine translation; model integration; bidirectional training strategy; mBART model