西北民族大学 2022 年全国机器翻译大会机器翻译评测技术报告

李亚超 *, 江涛, 加羊吉, 胡阿旭, 吕世良，祁坤钰

西北民族大学民族语言智能处理教育部重点实验室，甘肃省 兰州市，730030

摘要：本文介绍了西北民族大学参加的第十八届全国机器翻译大会中的藏汉、蒙汉以及维汉等三个翻译任务。由于藏语、维吾尔语测试集所包含的源语言输入句子较长，甚至是整篇文档，我们采用了针对性的处理方法，将较长句子切分成和训练语料长度符合的较短句子，并予以翻译。最后，我们提交了相应的翻译结果。

关键词：机器翻译、藏语、蒙古语、维吾尔语

中图分类号：TP391　　**文献标志码：** A

The Machine Translation Evaluation Technology Report on the 2022 China Conference on Machine Translation of Northwest Minzu University

LI Yachao, JIANG TAO, JIA Yangji, HU Axu, LV Shiliang, QI Kunyu

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou 730030, China

Abstract: This paper introduces the three translation tasks of Tibetan-Chinese, Mongolian-Chinese and Uyghur-Chinese on the 2022 China Conference on Machine Translation of the Northwest Minzu University. Because the source language input sentences contained in the Tibetan and Uyghur test sets are too long, even the whole document, we adopt a pre-processing method to segment the longer sentences into shorter sentences that match the length of the training corpus, and translate them. Finally, we submit the corresponding translation results.

Keywords: *Machine Translation; Tibetan; Mongolian; Uyghur*

1. 引言

本文介绍了西北民族大学参加的第十八届全国机器翻译大会（CCMT 2022）的相关情况和技术

* **通信作者：**harry_lyc@foxmail.com

要点。我们参加了藏语到汉语（藏汉）、蒙古语到汉语（蒙汉）、维吾尔语到汉语（维汉）三个翻译任务，基于 Transformer 模型[1]构建了一个神经机器翻译系统，并针对 CCMT2022 中藏语、蒙古语、维吾尔语的语料及语言特征，进行了相应的语料预处理，从而构建了藏语到汉语、蒙古语到汉语、维吾尔语到汉语等三个神经机器翻译系统。

我们参加的三个机器翻译任务，都采用受限测试，即严格利用主办方提供的训练语料，不利用额外的单语或双语数据。此外，不采用额外的其他数据，通过预训练或微调方式来提升模型性能。

2. 数据集

本文针对所有的语言，采用 SentencePiece 方法[2]进行分词，采用的词典大小为 30000。由于藏语、维吾尔语的测试集的输入句子较长，针对藏语测试语料，将源语言句子输入切分成不超过 32 个词语长度的子句子。针对维吾尔语测试语料，将源语言句子输入，将源语言句子输入切分成不超过 48 个词语长度的子句子。这对测试集句子切分的标准为：测试集语料的平均切分长度为其对应的训练集句子平均长度的 2 倍。经过实验表明，测试集句子切分长度为 2 倍的训练语料句子平均长度，能够取得最好的整体性能。

2.1 藏汉翻译数据集

藏汉翻译数据集由 CCMT2022 提供，约包含 115 万句对的训练语料。我们把 QHNU-test-tizh-CWMT2017 和 QHNU-test-tizh-CWMT2018 的数据集合并起来，作为测试集，包含 2778 个句子。同时把 2019TC_CCMT 和 QHNU-dev-tizh-CWMT2017 数据集合并起来作为开发集，包含 1646 个句子。

2.2 蒙汉翻译数据集

蒙汉翻译数据集由 CCMT2022 提供，约包含 126 万句对的训练语料。我们把 2019MC_test、IMU-test-mnzh-CWMT2017 和 IMU-test-mnzh-CWMT2018 的数据集合并起来，作为测试集，包含 3001 个句子。同时把 2020MC_CCMT 和 IMU-dev-mnzh-CWMT2017 数据集合并起来作为开发集，包含 2000 个句子。

2.3 维汉翻译数据集

维汉翻译数据集由 CCMT2022 提供，约包含 17 万句对的训练语料。我们把 DEV2019 和 XJIPC-test-uyzh-CWMT2018 的数据集合并起来，作为测试集，包含 1999 个句子。同时把 2020UC_CCMT 的数据集合并起来作为开发集，包含 1000 个句子。

3. 神经机器翻译模型

3.1 翻译模型

本文采用 Transformer 模型作为主要的翻译模型。Transformer 模型是深层的编码器-解码器结构。编码器和解码器的默认层数为 6 层，具体层数可以根据当前任务增加或减少。

编码器：编码器端包含 M 个相同的编码层，其中每一层包含了两个子层，分别为自注意力层和全连接的前馈神经网络层。为了缓解梯度问题，每个子层的输出都经过残差连接和层归一化处理。因为自注意力神经网络缺乏位置信息，需要在源语言词向量中加入显式的位置编码信息，通常采用正弦函数或可学习的位置编码方式实现。源语言位置编码信息输入到编码器的第一层中，然后依次经过各层编码，最后一层编码状态作为源语言表示。

解码器：解码器端包含 N 个相同的解码层，其中每一层包含了三个子层，分别为自注意力层、源语言注意力层和全连接的前馈神经网络层等。解码器端同样在每个子层应用残差连接和层归一化，目标语言词向量中同样加入位置编码信息。与编码器层相比，解码器端的每一层中增加了额外的注意力层用来获取源语言信息，也称为交叉注意力（Cross-Attention）。解码器的最后一层输出作为解码器的最终状态，输入 Softmax 层后，生成目标语言词语分布概率。

Transformer 模型自 2017 年提出来之后，在多个机器翻译任务上取得了最好的翻译效果，同时在其他自然语言处理任务上也获得了最佳性能。该模型成为当前主流的自然语言处理模型，得到广泛应用。本项目同样采用 Transformer 模型作为主要的翻译模型。

3.2 实验设置

本文系统均采用开源的 Transformer 模型实现 Fairseq[3]，具体采用大模型设置。在模型训练中，使用 Adam 优化方法[4]训练神经机器翻译模型，初始学习率为 0.0005，批量大小为 4096，更新频率为 4。采用学习率衰减策略，预热步骤均为 4000，标签平滑度和丢失率分别为 0.1。对于所有实验，采用共享双语词典的方法以降低计算量。

4. 实验结果

本节介绍本次评测的主要实验结果，采用 BLEU 评测译文质量。注意本文并不将汉语译文处理成字符序列，而是在词语序列上计算 BLEU 值。

表 1 主要实验结果

	藏汉翻译	蒙汉翻译	维汉翻译
翻译结果	45.13	47.43	24.63

从表 1 的实验结果可以看出，本文方法在所构建的测试集上取得了较好的实验结果。此外，本文同样采用一些其他的方法，比如 BPE 词语切分方法[5]，以及深层神经机器翻译模型等其他方法，

但是这些方法并没有能够显著提高翻译质量。因此，本文的主要实验结果，以及所提交的翻译译文，均采用 Transformer 大模型设置。

5. 结论

本文简单介绍了西北民族大学所参加的 CCMT2022 的藏汉、蒙汉、维汉等三个翻译任务。我们针对测试集和训练集的句子长度分布的不同特点，采用了针对性的预处理方法。此外，探索了一些不同的系统及翻译模型，来探索较好的模型设置，提交了相应的测试译文。

参考文献

[1] Vaswani A., Zhao Y., Fossum V., and Chiang D. Decoding with Large Scale Neural Language Models Improves Translation[A]. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing[C]. Seattle, USA: Association for Computational Linguistics, 2013:1387–1392.

[2] Kudo K., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations), 2018: pages 66–71.

[3] Ott M., Edunov S., Grangier D., Auli M. Scaling Neural Machine Translation[A]. In Proceedings of the Third Conference on Machine Translation[C]. Brussels, Belgium: Association for Computational Linguistics, 2018:1–9.

[4] Kingma, D.P., Ba, J. Adam: A Method for Stochastic Optimization[A]. In Proceedings of ICLR[C], 2015.

[5] Sennrich R., Haddow B., and Birch A. Neural Machine Translation of Rare Words with Subword Units[A]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics[C]. Berlin, Germany: Association for Computational Linguistics, 2016:1715–1725.