北京航空航天大学 CCMT2022 评测报告

唐子安，巢文涵*，徐贝凝

（北京航空航天大学，北京市 100191）

摘要：本报告介绍了本单位参加第十八届全国机器翻译大会(CCMT 2022)评测的情况。在本次评测中，我们参加了其中的 3 个翻译任务，分别是蒙汉综合领域机器翻译、藏汉综合领域机器翻译和维汉新闻领域机器翻译。以上翻译任务的主要问题为：维汉新闻领域语料资源稀缺。针对此问题，本系统采用了反向翻译方法提升翻译效果。实验表明，相对于基线系统，本系统采用的方法可以显著提升模型的翻译效果。

关键词：NMT（神经机器翻译）；BT（反向翻译方法）；Transformer

中图分类号：TP302.1　　**文献标志码：**A

1 引言

本报告介绍了本单位参加第十八届全国机器翻译大会(CCMT 2022)评测的情况。在本次评测中，我们参加了其中的 3 个翻译任务，分别是蒙汉综合领域机器翻译、藏汉综合领域机器翻译和维汉新闻领域机器翻译。

由于蒙汉和藏汉领域今年新增 100 万平行语句对，因此往年一直存在的低资源问题得到了很大程度的缓解，然而维语的平行语料依旧处于较少的状态，只有不到 30 万。考虑到本次评测提供了数百万的中文高质量单语语料，本次评测中我们针对维语模型的训练使用了 BT 的方式进行数据增强[1]。对于蒙语和藏语模型，由则直接使用给定的一百多万平行语料进行训练，视结果好坏再决定是否也使用 BT 方法。

在模型方面，我们使用了基于 mxnet-sockeye[2]实现的 Transformer[3]模型。

在验证集方面，因为维语方向的平行语料较少，我们采用了提供的 2020 年语料作为验证集。藏语和蒙语的语料资源相对丰富，因此我们直接从平行语料中抽取验证集。

2 数据

2.1 数据统计

我们参加的三个评测项目均提供了多组训练语料，包括双语平行语料和目标语言单语语料。对于双语语料，我们将其整合为一个文件用于模型的训练；对于单语语料，我们首先将文档中的句子进行提取，用于模型的反向翻译。

所有使用到的数据如表 1 所示（维语的括号内为 BT 语料数量）：

表 1 各方向语料数量统计

Tab.1 Data statistics of each evaluation corpus

	维-汉	藏-汉	蒙-汉
训练集	340114（170057）	1136578	1241454
验证集	1000	10000	10000
单语语料	5435051	5435051	5435051

*通信作者：chaowenhan@buaa.edu.cn

2.2 数据预处理

1、分词。对于汉语，我们使用了 jieba 对语料进行分词操作；因为 jieba 分词并不支持蒙藏维语，因此我们使用了 tokenizer 工具对这些语言进行了 tokenize 操作。

2、BPE（字节对编码）处理。我们使用 BPE 工具对平行语料进行了 BPE 的学习与处理，分别得到了汉蒙藏维的 BPE 词表。

3、生成 BPE factor。我们使用 factor 来区分一个 token 是否来自于一个更大的词汇。对于被 bpe 切开的词，我们对词中的 token 的 factor 赋予形如 BM（……）E 的值（若一个词被分为两个 token 则 factor 为 BE，4 个则为 BMME，以此类推），这些 factor 能够让模型更好地学习 token 与词的关系。

4、BT。由于维语资源稀少，因此我们利用 BT 对维语进行了数据增强。具体做法是先使用平行语料训练一反向（汉-维）模型，对汉语语料进行翻译后取翻译结果作为 BT 数据集，和原本的平行语料合并后作为最终的维汉训练语料。因为维汉原本的平行语料较少，增加太多合成语料可能会对训练结果造成不利影响，因此我们使用的平行语料和合成语料的比例为 1:1。

5、划分验证集。蒙汉与藏汉模型和维汉 baseline 模型在训练时使用的是从给定语料中抽取出的验证集，而维汉的最终模型使用了 CCMT2020 测试集作为验证集。

3 实验

3.1 实验环境

操作系统：Ubuntu 18.04.6
深度学习框架：mxnet 1.5.1
CPU：Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz
内存：256G
GPU：Tesla V100
显存：32G

3.2 实验设置

表 2 实验超参数设置

Tab.2 Hyperparameters setup

	蒙-汉	藏-汉	维-汉
SAN_dim	512	512	512
FFN_dim	2048	2048	2048
Multi_head	8	8	8
Enc_layer	6	6	6
Doc_layer	6	6	6
Dropout	0.1	0.1	0.1
Optimizer	Adam	Adam	Adam

3.3 实验过程

首先我们将所有语料进行了整合和简单的清洗。首先，我们使用给定的一百万语料训

练蒙汉、藏汉翻译模型，对翻译效果欠佳的蒙汉模型以及数据稀少的维汉模型，我们使用BT进行数组增强。

在BT过程中，我们使用提供的蒙汉和维汉语料训练了反方向的汉维和汉蒙模型，并尝试了BT。其中汉维模型的翻译效果比较让人满意，被我们采用。然而，汉蒙模型的训练结果显示，验证集bleu值异常偏低，经过数据核对、模型检查等一系列原因排查，仍未找到根本原因。因此，本次训练我们决定不对蒙汉语料进行数据增强。

在维汉BT完成后，我们将BT结果和原语料进行拼接，训练出了最终的维汉模型。

3.4 实验结果

因为离线评测的结果还未下发，以下结果均为验证集的结果：

表3 验证集实验结果

Tab.3 Results on Validation set

	蒙-汉	藏-汉	维-汉（采用BT）
bleu	35.56	53.22	27.86
chrf	48.44	62.36	38.22
Rogue1	42.89	59.37	55.13
RogueL	40.51	57.80	51.23

4 总结

相比于往年，本次CCMT2022的蒙语和藏语方向的语料资源充足，BT等其他数据增强方法在这种情况下增益不大；对于如维语这样的平行语料资源稀少的任务，BT效果更佳显著。遗憾的是，受限于时间因素，我们本次未能尝试其他如翻译后处理等提高准确率的方式，我们会在在线评测中再接再厉，继续研究和探索。

参考文献：

[1] Sennrich R , Haddow B , Birch A . Improving Neural Machine Translation Models with Monolingual Data[J]. Computer Science, 2015.

[2] Felix Hieber, Tobias Domhan, Michael Denkowski, et al . Sockeye: A Toolkit for Neural Machine Translation[J]. arXiv, 2017.

[3] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.

Beihang University CCMT2022 Report

Zian Tang, Wenhan Chao*, Beining Xu

（Beihang University，Beijing 100191）

Abstract： This report introduces our participation in the evaluation of the 18th National machine translation Conference (CCMT 2022). In this evaluation, we participated in three translation tasks, namely, Mongolian Chinese comprehensive domain machine translation, Tibetan Chinese comprehensive domain machine translation and Uygur Chinese news domain machine translation.

The main problem of the above translation task is: the corpus resources in the Uygur Chinese news field are scarce. To solve this problem, the system adopts the reverse translation method to improve the translation effect. The experiment shows that the method used in this system can significantly improve the translation effect of the model compared with the baseline system.
Keywords：NMT（Neural Machine Translation）；BT（Back Translation）；Transformer