北京理工大学 CCMT2021 技术报告

朱晓光，费伟伦，简林圳，鉴萍*，史树敏，武星

（北京理工大学，北京市 100081）

{xgzhu, wlfei, lzjian, pjian, bjssm, wuxing}@bit.edu.cn

摘要：本文详细介绍了本单位参加第十七届全国机器翻译大会(CCMT 2021)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，分别是蒙汉日常用语机器翻译、藏汉政府文献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：资源稀缺。针对此问题，本系统采用了掩码语言模型预训练、反向翻译等方法提升翻译效果。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：神经机器翻译；资源稀缺语言；Transformer 模型；预训练

中图分类号：TP302.1　　**文献标识码：**A

1 引言

本文详细介绍了本单位参加第十七届全国机器翻译大会（CWMT 2021）翻译评测任务的情况。我们共参与了 CCMT 2021 机器翻译评测任务中的三个有关少数民族语言的翻译项目，分别是蒙汉日常用语机器翻译、藏汉政府文献机器翻译和维汉新闻领域机器翻译。

相对于其他机器翻译任务，本次全国机器翻译大会提出的三种少数民族语言的机器翻译任务最显著的特征就是资源稀缺：三个翻译任务提供的双语平行句对数目均少于 30 万，均属于稀缺资源语言的机器翻译任务，增加了神经网络机器翻译模型的训练难度。

本次评测采用的系统和方法主要围绕资源稀缺问题进行设计，所以问题的关键就在于大规模单语语料的充分利用。我们参考 Sennrich 等人的工作[1]，利用评测提供的汉语单语语料通过反向翻译生成伪数据扩充神经机器翻译的训练数据。在此基础上，我们借鉴了 Caswell 等人的研究成果[4]，我们对扩展平行语料进行了特别处理，对于扩展平行语料中的源语言（少数民族语言）句子，我们在每句话的开头添加特别的标记“<BT>”。我们认为反向翻译同时引入了有益信号（引入了更多的知识）和有害信号（放大了机器翻译的偏差），向模型表明一个给定的训练句子是否被反译能够帮助模型区别有益信号和有害信号，提升模型的训练效果。特别地，针对维语-汉语翻译，我们利用掩码语言模型[5]将单语语料和反向翻译生成伪数据进行预训练，进一步提升了模型的表现。[1]

在实验中，我们采用谷歌 Transformer[3] 神经网络机器翻译架构作为基线翻译模型。除上文提到的方法外，我们对模型生成的译文进行了进一步的处理：我们通过对模型生成的翻译结果与参考结果进行对比，观察到部分翻译结果存在过翻译的问题，即存在多个重

基金项目：该工作受国家重点研发计划（2017YFB1002103）和国家自然科学基金（61732005）资助。

* 通信作者：pjian@bit.edu.cn

复的无意义字符串。针对这个问题，我们通过从长到短地枚举翻译结果句子的子字符串，删除相邻的重复子串，直到不存在相邻的重复子串。实验对比了翻译系统在三种少数民族语言翻译任务中不同实验设置下的表现，并对实验结果进行了分析。实验表明，我们的系统相对于基线系统在三种翻译任务上的译文质量均有明显的提升。

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，用于模型的预训练和反向翻译。

我们在参评系统中采用的训练数据统计如表 1 所示:

表 1 各评测语料数据统计

Table. 1 Data statistics of each evaluation corpus

	维语-汉语	藏语-汉语	蒙语-汉语
训练集	170, 061	157,959	262,643
验证集	1,000	1,000	1,001
测试集	1,000	1,000	2,001
单语语料	5,435,155	5,435,155	5,435,155

2.1 数据预处理

维语-汉语：

1. 分词。对于汉语，我们使用 jieba 对原始语料进行分词；对于维语，我们利用 Moses 工具集中的 tokenizer.perl 脚本对原始语料进行分词操作。

2. BPE 处理[2]。我们使用 fastBPE 工具对语料进行 BPE 的学习和处理，由于预训练模型的需要，我们将汉语和维语的语料混合后进行 BPE 的学习，其中 BPE CODES 大小设为 40000，学习完成后将学习到的 BPE 信息应用到平行语料中，最后抽取汉语和维语的联合词表，其长度为 48172。

3. 为了加快训练速度，我们将原始的文本信息进行索引化处理，即将文本中的词语替换成其在词表中的索引。

藏语-汉语：

1.由于平行语料的文件格式不一致，对所有文件进行平行语料的提取，生成格式一致的文件，在这个过程中，对汉语和藏语的句子进行处理，如全角半角转换，非法字符删除

等；

2.由于所给的数据并没有明确划分验证集、测试集，根据训练及测试需要，自行划分测试集与验证集。划分完毕后，得到验证集 1379 条，测试集 1000 条；

3.为提升模型的泛化能力，将训练集中与验证集和测试集重复的句对删除，其中训练集与验证集的重复句对 18 条，与测试集的重复句对 126 条；

4.本文的汉语分词工具采用 jieba 分词。在测试过程中我们也曾使用现有的藏语分词工具对藏语进行分词，但由于效果不如按音节的效果好，因此未对藏语再做分词处理；

5.采用 Moses 中的 normalize-punctuation.perl、tokenizer.perl 脚本，对两种语言的数据进行标点符号标准化和进一步切分；

6.对以词为单位分割的语料，采用 subword-nmt 进行训练 BPE 并应用于语料，BPE 操作符规模设置为藏语 10k、汉语 20k；

7.采用 Moses 中的 clean-corpus-n.perl 脚本，过滤句子过长或双语句子长度比过大或过小的句对；

8.从中文单语数据中根据句子长度选取大约三百万条数据，利用反向翻译对语料进行扩展。

经过上述处理后，我们获得的数据规模如下：

过滤后的训练集：155082

扩展后的训练集：3895088

验证集：1287

测试集：1000

蒙语-汉语：

1. 过滤平行语料中含有乱码的句对，主要包括源语言端（少数民族语言）包含汉语的句对以及目标语言端（汉语）包含少数民族语言的句对；

2. 特别的，1001 条验证集语料中有 294 条数据被包含在训练集数据中，为了客观地展示翻译模型的泛化能力将之从训练集中删去；

3. 采用 hanlp 对汉语进行中文分词；

4. 采用 Moses 中的 normalize-punctuation.perl、tokenizer.perl 脚本，对两种语言的数据进行标点符号标准化和进一步切分；

5. 对以词为单位分割的语料，采用 subword-nmt 进行训练 BPE 并应用于语料，BPE 操作符规模设置为蒙语 24k、汉语 16k；

6. 采用 Moses 中的 clean-corpus-n.perl 脚本，过滤句子过长或双语句子长度比过大或过小的句对；

7. 从中文单语数据中选取大约三百万条数据，利用反向翻译对语料进行扩展。

经过以上步骤处理后得到的训练语料统计信息为：

过滤后训练集：237979

扩展后训练集：3284127

验证集：1001

3 方法

3.1 维语-汉语

在维语-汉语的翻译中，我们的翻译模型分为四个阶段：

1. 汉语-维语翻译模型训练。我们首先利用现有的平行语料训练汉语-维语的翻译模型，然后利用收敛后的模型对现有的汉语单语语料进行翻译，从而得到反向翻译的扩展语料。该扩展语料将被用作预训练和翻译模型的训练。

2. 预训练阶段。我们采用掩码语言模型（MLM）进行预训练，采用的模型为 Transformer 的编码器部分，训练所用的数据为汉语单语语料以及阶段一得到的反向翻译的扩展语料。

3. 维语-汉语翻译模型训练。首先，我们将预训练阶段学习到的模型参数用来初始化翻译模型的编码器和解码器部分。特别地，对于解码器中存在特有的 Encoder-Decoder Attention 组件，我们采用随机初始化的处理方式。然后，我们将阶段一得到的扩展平行语料（随机采样 300 万）与人工标注的语料（过采样处理）进行混合，训练最终的维语-汉语翻译模型，直至收敛。特别需要指出，我们对扩展平行语料进行了特别处理，对于扩展平行语料中的维语句子，我们在每句话的开头添加特别的标记“<BT>”。我们认为反向翻译同时引入了有益信号（引入了更多的知识）和有害信号（放大了机器翻译的偏差），向模型表明一个给定的训练句子是否被反译能够帮助模型区别有益信号和有害信号，提升模型的训练效果。

4. 译文后处理。我们通过对模型生成的翻译结果与参考结果进行对比，观察到部分翻译结果存在过翻译的问题，即存在多个重复的无意义字符串。针对这个问题，我们通过从长到短地枚举翻译结果句子的子字符串，删除相邻的重复子串，直到不存在相邻的重复子串。

3.2 藏语-汉语

在藏语-汉语的翻译中，我们的翻译模型共分为四个部分：

1. 汉语-藏语翻译模型训练。利用已有的藏汉平行语料训练一个汉语-藏语的神经机器翻译模型，直到该模型收敛。用训练好的汉语-藏语翻译模型将筛选后得到的汉语语料做翻译，得到大量的伪语料，用于之后的藏语-汉语翻译模型的训练。

2. 尽管我们利用汉语单语语料得到了大量的伪语料，但是由于该语料是由翻译模型得到的，与真实的平行语料相比，伪语料的数据分布可能更加单一。因此我们需要帮助模型分辨有哪些数据是真实的平行语料数据，而哪一部分和真实的分布是不同的。借鉴前人的工作，我们对数据又做了新一步的处理：在伪语料中的藏语句子开始部分添加标记“<BT>”，表示该条数据是由翻译模型生成的，也可以认为该条数据的分布与真实的数据分布具有差异性，方便模型从混合的数据中学到真正有用的信息和内容。

3. 考虑到我们得到的汉语单语语料领域是新闻领域的，而本次藏汉翻译的数据均来

自政府公文。如果直接使用混合数据（伪语料与真实的平行语料组成）训练得到的数据做测试，可能会使领域信息丢失。因此我们在完成第 2 步的训练之后使用真实的平行语料又做了一次类似微调的工作，不仅为了使模型学到的分布更加趋向于真实的藏汉语料分布，也可以让模型学到更多关于领域的信息。

4. 译文后处理。主要包括了全角半角转换，重复字符的删除等。

3.3 蒙语-汉语

在蒙语-汉语的翻译中，我们的翻译模型分为三个阶段：

1. 汉语-蒙语翻译模型训练。我们首先利用现有的平行语料训练汉语-蒙语的翻译模型，然后利用收敛后的模型对现有的汉语单语语料进行翻译，从而得到反向翻译的扩展语料。该扩展语料将被用作预训练和翻译模型的训练。

2. 蒙语-汉语翻译模型训练。我们将阶段一得到的扩展平行语料（随机采样 300 万）与人工标注的语料（过采样处理）进行混合，训练蒙语-汉语翻译模型，直至收敛。特别地，我们在扩展平行语料源句子的开头添加特别的标记“<BT>”。

3. 译文后处理。同维汉翻译。

4 实验

4.1 实验环境

维语-汉语：

操作系统：CentOS Linux release 7.9.2009 (Core)
深度学习框架：Pytorch 1.0
CPU：AMD EPYC 7742 64-Core Processor
内存：200 GB
GPU：GTX 1080
显存：11GB

藏语-汉语、蒙语-汉语：

操作系统：CentOS Linux release 7.9.2009 (Core)
深度学习框架：Pytorch 1.7.0
机器翻译框架：fairseq 0.10.2
CPU：AMD EPYC 7742 64-Core Processor
内存：200 GB
GPU：GeForce RTX 3090
显存：24GB

4.2 实验设置

表 2 实验设置信息

Table. 2 Experimental setup

	维语-汉语	藏语-汉语	蒙语-汉语
Emb_dim	1024	1024	1024
FFN_dim	4096	4096	4096
编码器层数	6	6	6
解码器层数	6	6	6
Dropout	0.3	0.3	0.3
Label_smoothing	0.1	0.1	0.1
Optimizer	Adam	Adam	Adam

4.3 实验结果及分析

维语-汉语：以下实验的评估指标均为 BLEU4，以单个汉字或符号作为评估的基本单位。维-汉翻译任务中，测试集、开发集与训练集均为新闻领域，与单语语料领域相同，因此，相对于蒙-汉、藏-汉翻译任务更加适合采用语料扩展方法提升翻译效果。实验结果表明，采用利用预训练+单语语料反向翻译方法译文质量提升明显（+6.8 BLEU4）。由此可见，领域适应问题对于机器翻译十分重要。

表 3 维语-汉语 BLEU 结果，beam_size=8

Table 3. The BLEU score on uy-zh, beam_size=8

编号	设置	验证集	测试集
1	基线模型	46.02	38.33
2	预训练	47.91(+1.89)	40.97(+2.64)
3	预训练+标签的反向翻译	**52.63(+6.61)**	**45.13(+6.8)**

藏语-汉语：以下实验的评估指标均为 BLEU4，以单个汉字或符号作为评估的基本单位。根据划分出来的测试集上的实验结果，我们发现，在基线模型的基础上，隐层单元更大的 Transformer 大参数版本可以有效地提升翻译效果；反向翻译也有所提升；在此基础上加入带标签的反向翻译，对效果的提升也有帮助；经过大量数据训练后的模型，再由真实的平行语料进行微调，提升幅度相对更大；后处理的提升相对比较有限。

表 4 藏语-汉语BLEU结果，beam_size=8

Table 4. The BLEU score on ti-zh, beam_size=8

编号	设置	测试集
1	基线模型	45.29
2	+Transformer 大参数版本	45.67(+0.38)
3	+带标签的反向翻译	45.88(+0.21)
4	+双语语料微调	46.25(+0.96)
5	+后处理	46.37(+1.08)

蒙语-汉语：以下实验的评估指标均为 BLEU4，以单个汉字或符号作为评估的基本单位。根据验证集上的实验结果，我们发现，在基线模型的基础上，隐层单元更大的 Transformer 大参数版本可以有效地提升翻译效果；反向翻译的增益最大，可以显著地提升翻译效果；在此基础上，带标签的反向翻译对于翻译效果也有较大提升；后处理对于翻译效果仅有较小的提升。

表 5 蒙语-汉语 BLEU 结果，beam_size=14

Table 5. The BLEU score on mn-zh, beam_size=14

编号	设置	蒙-汉
1	基线模型	41.30
2	+Transformer 大参数版本	42.53(+1.23)
3	+反向翻译	45.09(+3.79)
4	+带标签的反向翻译	45.86(+4.56)
5	+后处理	**45.95(+4.65)**

5 总结

在本次评测中，我们参与了三个少数民族语言到汉语的翻译评测项目。三种翻译任务所共有的问题是资源稀缺。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 BPE、掩码语言模型、反向翻译、译文后处理等方法应对上述三个问题，提升翻译质量。实验表明我们采用的方法相对于基线系统对翻译质量有所提升，证明其可以有效地缓解上述问题对翻译结果带来的负面影响。

限于时间上的限制，我们所提的方法还有一定的提升空间。未来，我们将在模型集成、数据筛选等方向作进一步的研究。

6 参考文献

[1] Sennrich R , Haddow B , Birch A . Improving Neural Machine Translation Models with Monolingual Data[J]. Computer Science, 2015.
[2] Sennrich R , Haddow B , Birch A . Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.
[3] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.
[4] Caswell I , Chelba C , Grangier D . Tagged Back-Translation[J]. 2019.
[5]BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

Beijing Institute of Technology CCMT2021 Report

Xiaoguang Zhu, Weilun Fei, Linzhen Jian, Ping Jian*, Shumin Shi,

Xing Wu

(Beijing Institute of Technology, Beijing 100081)
{xgzhu, wlfei, lzjian, pjian, bjssm, wuxing}@bit.edu.cn

Abstract：This paper introduces in detail the evaluation of our unit participating in the 17th National Machine Translation Conference (CCMT 2021). We participated in three translation tasks, namely, Mongolian-Chinese daily language machine translation, Tibetan-Chinese government document machine translation and Uygur-Chinese news machine translation. The main problem of the above translation tasks is that resources are scarce. In order to solve this problem, this system adopts the method of mask language model pre-training and back-translation to improve the translation effect. Experiments show that compared with the baseline system, the method adopted by the system can significantly improve the translation effect of the model.

Keywords：Neural Machine Translation；Low Resource Language；Transformer Model；Masked Language Model