北京理工大学 CCMT2022 技术报告[1]

依西降参，简林圳，朱晓光，史树敏*，鉴萍*

（北京理工大学计算机学院，北京市，100081）

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

摘要：本文详细介绍了北京理工大学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，分别是维汉新闻领域机器翻译、蒙汉日常用语机器翻译和藏汉政府文献领域机器翻译。本系统分别采用了掩码语言模型预训练、反向翻译、滑动参数平均、集成学习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；Transformer 模型；预训练；反向翻译；集成学习

中图分类号：TP302.1　　**文献标志码：** A

1 引言

本文介绍了本单位参加第十八届全国机器翻译大会（CCMT 2022）机器翻译评测任务的情况。我们共参与了 CCMT 2022 机器翻译评测任务中的 3 个翻译项目，分别是维汉新闻领域机器翻译、蒙汉日常用语机器翻译和藏汉政府文献领域机器翻译。

我们所参加的 3 个机器翻译任务均在低资源下进行，而基于神经网络的机器翻译系统甚受平行语料数量的影响，因此为了解决平行语料资源限制，我们通过使用反向翻译[1]手段来试图解决该问题，同时为了提高模型的鲁棒性，我们也使用了集成学习[2]和掩码语言模型预训练[3]等方法来提高机器翻译的翻译质量。

在实验中，我们采用谷歌 Transformer [4]神经网络机器翻译架构和 Wu[5] 等提出的 LightConv 神经网络机器翻译架构作为基线翻译模型。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来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[6]，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-汉语翻译任务，我们利用掩码语言模型[3]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。针对藏语-汉语翻译任务，我们利用参数平均和模型融合等集成学习[2]方法来进一步提高翻译效果。实验表明，我们的系统相比于基线系统在维语-汉语、蒙语-汉语和藏语-汉语等三种翻译任务上的译文质量均有明显的提高。

2 数据

2.1 数据预处理

[1] **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：**bjssm@bit.edu.cn

维语-汉语：

1. 分词。对于汉语，我们使用 jieba[2] 对原始语料进行分词；对于维语，我们利用 Moses[3] 工具集中的 tokenizer.perl 脚本对原始语料进行分词操作。

2. BPE 处理[7]。我们使用 fastBPE [4] 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。

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（蒙语）包含汉语的句对以及目标语言端（汉语）包含蒙语的句对；

2. 去掉训练集数据中心的重复句对；特别的，验证集语料中有部分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[5] 分词工具对汉语数据进行分词；

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[6] 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。

藏语- 汉语:

1. 分词：采用 jieba 分词工具对汉语数据进行分词，用 TIP-LAS [8] 分词工具对藏文数据进行分词；

2. 采用 Moses 中的 normalize-punctuation.perl 脚本，对两种语言的数据进行标点符号标准化。

3. BPE 处理：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 25k、汉语 25k。

4. 从中文单语数据中选取大约一百万条数据，利用反向翻译对语料进行扩展。

我们将主办方提供的双语平行语料数据和目标语言单语数据进行整理后的得到的数据条数如表 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	蒙语-汉语	藏语-汉语
训练集	170,061	1,189,981	1,156,580
训练集	1,000	2,001	2,049
测试集	1,000	2,001	2,379
单语语料	5,435,155	5,435,155	5,435,155

[2] https://github.com/fxsjy/jieba.git
[3] https://github.com/moses-smt/mosesdecoder.git
[4] https://github.com/glample/fastBPE.git
[5] https://github.com/hankcs/HanLP.git
[6] https://github.com/rsennrich/subword-nmt.git

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三个阶段：

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。

3.2 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。我们首先利用初始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。

2. 蒙语-汉语翻译模型训练。我们将初始的平行语料（过采样处理）与阶段一得到的扩展平行语料进行混合，训练蒙语-汉语翻译模型。特别地，我们在扩展平行语料源语言句子的开头添加标记“<BT>”。

3.蒙语-汉语翻译模型微调。我们将阶段二训练好的蒙汉翻译模型再经过初始的平行语料进行微调，直至收敛。

3.3 藏语- 汉语

在藏语-汉语的翻译中，我们的翻译模型的集成学习分为三个阶段：

1. Transformer 藏汉翻译模型训练：我们利用官方提供的平行语料训练基于 Transformer 的藏汉翻译模型，模型收敛后对最后 5 个检查点的参数矩阵进行平均并保存与评估。

2. Lightconv 藏汉翻译模型训练：我们利用官方提供的平行语料训练基于 Lightconv 的藏语-汉语翻译模型，模型收敛后对最后 5 个检查点的参数矩阵进行平均并保存与评估。

3.融合不同藏汉翻译模型：我们在解码过程中经过 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，进而预测下一个词来试图提高藏汉翻译效果。

我们的藏汉翻译模型的反向翻译也分为三个阶段：

1. Transformer 汉藏翻译模型训练：我们首先利用真实的平行语料训练汉语-藏语的翻译模型，然后利用收敛后的模型对现有的汉语单语语料进行翻译，从而得到伪平行语料。在生成伪平行语料时，根据波束搜索得到 k=7 个最可能的目标语句后，选取其中概率最高的作为目标语句；
2. 为了跟上一步中生成伪平行语料的方法作对比，我们参考 Edunov S[9]的工作，根据波束搜索得到 k=10 个最可能的目标语句后，随机采样其中任一语句作为目标语句；
3. Transformer 藏汉翻译模型训练：根据上两步不同的伪平行语料生成方法，我们将真实的平行语料（过采样处理）与不同的生成方法得到的伪平行语料分别进行混合，分别训练不同的藏语-汉语翻译模型。

4 实验

4.1 实验环境

操作系统：CentOS Linux release 7.9.2009 (Core)
深度学习框架：Pytorch 1.7.1（蒙语-汉语）、Pytorch 1.0（维语-汉语）和 Pytorch 1.11.0（藏语-汉语）。
机器翻译框架：fairseq 0.10.2（蒙语-汉语）和和 fairseq 0.12.0（藏语-汉语）。
CPU：AMD EPYC 7742 64-Core Processor
内存：200 GB
GPU：GTX 1080
显存：11GB

4.2 实验设置

表 2 实验设置信息

Table. 2 Experimental setup

	维语-汉语	蒙语-汉语	藏语-汉语
Emb_dim	1024	512	512
FFN_dim	4096	2048	2048
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 实验设置信息

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

编号	设置	验证集(BLUE4)	测试集(BLUE4)
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)

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，反向翻译能够带来较大的收益；在此基础上，滑动参数平均带来的提升基本可以忽略。

表 4 实验设置信息

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

编号	设置	验证集(BLUE4)	测试集(BLUE4)
1	基线模型	56.47	54.42
2	+带标签的反向翻译	57.46(+0.99)	55.14(+0.72)
3	+滑动参数平均	57.48(+1.01)	55.15(+0.73)

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。根据验证集上的实验结果，我们发现，模型的集成学习和反向翻译能够带来较好的收益。

表 5 实验设置信息

Table. 5 The BLEU score on ti-zh, beam_size=7

编号	设置	测试集（BLUE4）
1	Transformer	48.47
2	LightConv	47.98
3	1+参数平均	49.64(+1.17)
4	2+参数平均	49.60(+1.62)
5	1 和 2 模型融合	50.68(+2.21)
6	3 和 4 模型融合	51.25(+2.78)
7	1+反向翻译（beam search）	48.92（+0.45）
8	1+反向翻译(top10)	49.96(+1.49)

5 总结

在本次评测中，我们参与了 3 个翻译评测项目。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、反向翻译和集成学习等方法提升翻译质量，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。

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] Tong X , Zhu J , Liu T . Bagging and Boosting statistical machine translation systems[J]. Artificial intelligence, 2013, 195(FEB.):496-527.
[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[4] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.
[5]Wu F, Fan A, Baevski A, et al. Pay less attention with lightweight and dynamic convolutions[J]. arXiv preprint arXiv:1901.10430, 2019.
[6] Caswell I , Chelba C , Grangier D . Tagged Back-Translation[J]. 2019.
[7] Sennrich R , Haddow B , Birch A . Neural Machine Translation of Rare Words with Subword Units[J]. Computer Science, 2015.
[8] 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.
[9] Edunov S , Ott M , Auli M , et al. Understanding Back-Translation at Scale[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

Beijing Institute of Technology CCMT2022 Report

Jiangcan Yixi, Linzhen Jian,Xiaoguang Zhu, Ping Jian *, Shumin Shi *

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081,China)

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

Abstract: This paper introduces in detail the evaluation of our unit participating in the 18th National Machine Translation Conference (CCMT 2022). We participated in three translation tasks, namely, Uygur-Chinese news machine translation，Mongolian-Chinese daily language machine translation and Tibetan-Chinese government document 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, back-translation and ensemble learning 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; Ensemble Learning