中科院计算所 CCMT 2022 蒙汉翻译技术报告

房庆凯[12†]，马铮睿[12†]，桂尚彤[12†]，张倬诚[12]，伍烜甫[12]，黄浪林[12]，

张民[3]，冯洋[12*]

（1. 中国科学院计算技术研究所，北京 100190；2. 中国科学院大学，北京 100049；3. 哈尔滨工业大学（深圳），深圳 518055）

摘要：本文介绍了中国科学院计算技术研究所参加第十八届全国机器翻译大会（CCMT 2022）蒙汉机器翻译评测的技术实现。本次参评系统使用深层 Transformer 作为基线模型，通过启用 dropout 的反向翻译进行数据增广，结合真实平行语料和伪平行语料一起训练模型。在此基础上，引入正则化方法 PD-R 辅助模型训练。在解码阶段，将不同数据训练得到的模型的神经网络输出层概率进行平均。实验结果表明，本文的方法在翻译质量上相比基线系统取得了显著的提升。

关键词：神经机器翻译；反向翻译；正则化

中图分类号：TP391.2　　**文献标志码：** A

1 引言

机器翻译旨在利用计算机将一种语言下的文字翻译到另外一种语言，有着极为广阔的应用场景。近年来，随着深度学习技术的进步，神经机器翻译取得了飞速的发展。目前，基于 Transformer[1]的神经机器翻译模型已经能够产生十分流畅且高质量的译文。第十八届全国机器翻译大会（CCMT2022）开展了多项翻译任务的评测，本文详细介绍了本单位参加蒙汉机器翻译评测任务的技术方案。

在本次评测中，我们针对蒙汉机器翻译任务，首先构建了基于深层 Transformer 的基线模型，相比普通的 Transformer 模型有更高的性能。为了扩大训练数据的规模，本系统基于反向翻译[10]技术，利用额外的中文语料制造伪平行数据。过滤掉质量较差的伪平行数据后，将其与真实数据结合到一起训练，相比基线模型能够取得显著的提升。为了获取更加多样的伪数据，在反向翻译过程中我们启用了 dropout[3]，通过设置不同的随机种子来得到多份不同的伪数据。此外，为了更好的优化模型，防止模型对样本产生过拟合或欠拟合，本系统利用正则化技术 Prediction Difference Regularization (PD-R)[2]辅助模型训练，使模型性能能够进一步提升。最后，将利用不同伪数据训练得到的模型进行集成，解码得到译文。实验结果表明，本文提出的方法能够在强基线系统之上带来超过 5.0 个 BLEU 分数的提升。

2 数据处理

2.1 预处理

本次评测组织方同时提供了蒙汉平行语料以及中文的单语语料。蒙汉平行语料的训练集由若干数据集组合而成，共包含约 126 万条平行句对；验证集包含约 2000 条平行句对。平行语料的预处理步骤如下：

1. 我们发现许多句对同时出现在了训练集和验证集中，为了保证验证集的合理性，我们将这些句对从训练集中去除；
2. 去除数据集中出现多次的句对；

基金项目：国家重点研发计划项目（NO. 2017YFE0192900）

† **相同贡献**

* **通信作者：**fengyang@ict.ac.cn

3. 去除训练集中语言错误的句对，即源端非蒙语或目标端非中文的句对；
4. 使用 moses 对中文的标点符号进行正规化；
5. 使用 jieba 对中文进行分词，蒙语不做分词；
6. 对蒙语和中文分别做 32K 次 BPE[6]合并操作，并对中文的词表进行截断，最终蒙语和中文词表大小分别为 32626 和 40000，具体见 4.1 节。

最终得到约 125 万平行句对。

中文单语语料包含若干网上爬取的新闻，预处理步骤如下：

1. 使用 moses 对标点符号进行正规化；
2. 通过正则表达式分别提取新闻的标题和正文，对二者分别进行切分，得到句子级的单语语料；
3. 去除数据集中出现多次的句子；
4. 使用 jieba 进行分词；
5. 去除长度过长的句子；
6. 去除所有包含英文字符的句子；
7. 使用与处理平行语料时相同的 BPE 词表，对句子进行 BPE 切分，具体见 4.1 节。

最终得到约 480 万条中文句子。

2.2 数据增广

为了充分利用中文单语数据，扩大蒙汉训练数据的规模，我们采用反向翻译[10]技术来进行数据增广。具体来说，我们利用蒙汉平行数据训练了一个中文→蒙语的基线系统（具体配置见 3.1 节），为 2.1 节中处理后的中文单语语料生成对应的蒙语伪翻译。在反向翻译解码时，我们使用 beam search 算法获得得分最高的译文。

由于中文单语语料的领域与平行语料存在不一致的情况，反向翻译产生译文的质量无法完全得到保证。为了防止低质量伪数据对模型训练带来负面影响，我们对反向翻译得到的伪数据进行了进一步过滤，去除质量较低的平行句对。

为了进一步增强伪数据的多样性，我们在反向翻译过程中启用了模型的 dropout 模块，为模型引入随机性。通过设置不同的随机种子，可以得到多份不同的伪平行语料。对于每份平行语料，都按照上述步骤进行过滤。

最终，我们通过不启用 dropout 的反向翻译得到 1 份伪平行语料，通过启用 dropout、设置不同随机种子的反向翻译得到 4 份伪平行语料。数据集的完整统计信息见表 1。

表 1 数据集统计信息

Tab.1 Statistics of the dataset

	真实数据	反向翻译	启用 dropout 的反向翻译			
			随机种子 1	随机种子 2	随机种子 3	随机种子 4
训练集	1250533	4236382	4135208	4135455	4135267	4135876
验证集	1998	-	-	-	-	-

3 方法

本次评测我们采用了基于自注意力机制的深层 Transformer[1]模型作为基线系统。在此基础上，我们采用了 Prediction Difference Regularization (PD-R)[2]正则化方法，防止模型对样本过拟合或欠拟合。最终，将基于不同数据训练得到的模型的神经网络输出层概率集成进行解码，得到最终的译文。

3. 1 基线系统

本次评测的基线系统基于 Transformer，首先，我们探索了不同配置的 Transformer，确保基线模型的性能足够强。具体来说，我们尝试了 Transformer-Base、Transformer-Base-Deep、Transformer-Big-Deep 三种配置的模型，其具体参数配置见表 2。

表 2 Transformer 的三种模型配置

Tab.2 Three configurations of Transformer

配置	Base	Base-Deep	Big-Deep
encoder_layers	6	15	15
decoder_layers	6	6	6
embed_dim	512	512	1024
ffn_embed_dim	2048	2048	4096
attention_heads	8	8	16
Dropout	0.1	0.1	0.3

在蒙汉真实平行语料上分别训练三种配置的模型，验证集上的性能见表 3。最终，我们选择 Transformer-Big-Deep 配置作为我们的基线系统。

表 3 不同配置 Transformer 的性能对比

Tab.3 Results on development set under different configurations.

配置	BLEU
Base	54.73
Base-Deep	54.42
Big-Deep	**57.36**

3. 2 对抗噪声的预测差异正则化

在机器翻译模型训练过程中，在输入端添加噪声被认为是有利于提升模型的泛化性能[4]。近期的工作 PD-R[2]同时表明，基于对抗噪声的输入和原始输入差异的正则化训练方式，能够避免模型对样本过拟合或欠拟合，进一步提升翻译模型在测试集上的性能。

首先介绍 PD-R 方法：对于训练集中的一条平行数据(X, Y)，其中X为源端句子，Y为目标端句子。为源端和目标端句子分别加噪声得到扰动后的样本$(\widetilde{X}, \widetilde{Y})$，分别将$(X, Y)$和$(\widetilde{X}, \widetilde{Y})$输入网络，对解码器输出的概率分布进行正则化，正则化损失可以表示为：

$$\mathcal{L}_{PD-R}(X, Y, \theta) = R\left[P(* \mid X, Y_{<}, \theta'), P\left(* \mid \widetilde{X}, \widetilde{Y}_{<}, \theta''\right)\right] \quad (1)$$

其中，θ表示模型参数，θ'和θ''分别表示由不同 dropout 得到的不同子网络，R代表某种衡量分布间距离的函数。

最终训练损失函数为：

$$\mathcal{L} = \mathcal{L}_{CE}(D,\theta) + \gamma\mathcal{L}_{PD-R}(D,\theta) \quad (2)$$

其中，$\mathcal{L}_{CE}$为机器翻译中常用的交叉熵损失函数。

3. 3 神经网络集成解码

在 2.2 节中，我们通过启用 dropout 的反向翻译技术得到了多份伪平行数据。通常，在不同数据集上训练得到的模型集成后会带来性能的提升。我们按照上述方法在不同数据集上训练了多个模型，在解码阶段对多个模型进行 token 级别的集成解码。4.2 节中的实验结果表明，集成解码能够有效提升模型在测试集上的性能。

4 实验

4. 1 实验设置

我们使用了基于 *PyTorch* 的开源工具 *Fairseq*[5]来实现我们的机器翻译系统。为了使模型对未登录词鲁棒，我们使用了开源工具 subword-nmt 分别对蒙文语料和中文语料进行了 32K 次 BPE[6]合并操作，得到了大小分别为 32626 和 47047 的蒙文和中文词表。为了防止词表过大，我们对中文词表进一步采取了截断操作，将词表大小限制在 40000。我们使用 4 张 RTX 3090 GPU 训练模型。

在双语语料上训练基线模型时，我们使用了 Adam[7]作为优化器，训练过程中 dropout[3]和 label smoothing[8]的值分别被设置为 0.3 和 0.1，其余训练参数设定与 *Fairseq* 的默认设定相同。为了加快模型训练速度，我们使用了混合精度训练以及分布式训练，训练中每一步的 batch 大小被设置为 32768 tokens。

在利用单语中文数据制造伪平行语料时，我们将 beam 大小和长度惩罚分别设置为 4 和 0.6。为了制造源端的多样性，我们启用了模型的 dropout 模块并使用了不同的随机种子额外进行了 4 次解码，总共生成了 5 份不同的伪平行语料。

我们将生成的每一份伪平行语料与真实平行语料分别结合后从头开始训练我们的模型。训练结束后，我们均挑选了在验证集上性能最好的 5 个 checkpoint 进行参数平均。

在最终解码时，我们将基线模型与 5 个不同的反向翻译增强的模型的神经网络输出层概率进行了平均。beam 大小和长度惩罚分别设置为 5 和 0.6。

4. 2 实验结果

我们将评测委员会所发布的 2019 和 2020 年的验证集合并作为我们的验证集，并报告在验证集上的结果。合并后的验证集一共有 1998 条蒙汉双语对。我们对模型的中文输出进行去除 BPE 和去除空格操作，然后使用 SacreBLEU[9]计算 detokenized BLEU。

表 4 显示了不同增强策略在验证集上的结果。所有的单语反向翻译策略相比于基线模型均能提升翻译质量。其中，标准的反向翻译策略能带来约 2.77 个 BLEU 分数的提升；而启用模型 dropout 模块的反向翻译策略平均能带来约 3.13 个 BLEU 分数的提升，说明了向反向翻译构造的伪平行语料的源端适度添加噪音与多样性能够进一步提升模型性能。其次，在训练过程中采用正则化方法平均能够进一步提升约 0.65 个 BLEU 分数，说明了在使用更深的模型架构时正则化对性能的重要性。

表 4 不同增强策略对翻译质量的影响

Tab.4 Results of the Mongolian-Chinese translation on development set under different enhancement strategies

	基线系统	+反向翻译	+启用 dropout 的反向翻译			
			随机种子 1	随机种子 2	随机种子 3	随机种子 4
无正则化	57.36	60.13	**60.82**	60.31	60.28	60.52
PD-R[2]正则化	**57.85**	**61.17**	60.78	**61.21**	**61.08**	**61.40**

我们使用表 4 的六种不同单语增强策略得到的模型（包括基线模型）作为基础模型进行神经网络输出层集成。每种单语增强策略根据其在验证集上的表现决定是否使用正则化技术。在集成时，集成模型的概率分布输出由基础模型的概率分布输出平均得到。表 5 的前 6 行显示了不同基础模型在验证集上的性能，后 4 行显示了不同集成策略得到的模型在验证集上的性能。我们有如下发现：

表 5 不同集成策略对翻译质量的影响

Tab.5 Results of the Mongolian-Chinese translation on development set under different ensemble strategies

序号	方法	BLEU
1	基线系统	57.85
2	+反向翻译	61.17
3	+反向翻译（dropout#1）	60.82
4	+反向翻译（dropout#2）	61.21
5	+反向翻译（dropout#3）	61.08
6	+反向翻译（dropout#4）	61.40
7	4+6	61.94
8	2+4+6	62.14
9	1+2+4+6	62.37
10	1+2+3+4+5+6	**62.54**

1） 对由不同随机种子得到的启用模型 dropout 模块的反向翻译语料所训练的模型进行集成，可以提升模型性能（第 4、6 行 vs 第 7 行）。这显示了启用模型 dropout 模块的反向翻译技术带来的源端多样性的好处。

2） 在第 7 行的基础上，进一步集成普通反向翻译技术得到的语料所训练的模型可以进一步提升翻译质量（第 8 行 vs 第 7 行）。

3） 在第 8 行的基础上，进一步集成只在真实平行语料上所训练的模型可以进一步提升翻译质量（第 9 行 vs 第 8 行）。

我们最后提交的结果由集成所有 6 个基础模型得到（第 10 行）。相比于普通基线模型，我们在验证集上得到了 5.18 个 BLEU 得分的提升。在在线测试平台的 CCMT2021 测试集上，我们的模型取得了 56.49 的 BLEU-SBP 得分，相比于普通基线模型提高了 6.66 个 BLEU-SBP 得分。

5 总结

本文介绍了中科院计算所参加 CCMT2022 蒙汉机器翻译评测任务的技术方案。本系统基于深层 Transformer 的强基线模型，通过启用 dropout 的反向翻译生成具有多样性的高质量伪平行语料。将真实语料与伪平行语料结合训练，并引入正则化方法防止模型对数据过拟合或欠拟合，最终集成在不同数据上训练的多个模型进行解码。在本地验证集及在线测试集上的结果均表明，本文提出的一系列方法相比强基线模型能够取得显著的提升。

参考文献：

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Guo, D., Ma, Z., Zhang, M., & Feng, Y. (2022, May). Prediction Difference Regularization against Perturbation for Neural Machine Translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 7665-7675).

[3] Srivastava N， Hinton G， Krizhevsky A , et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[4] Takase, S., & Kiyono, S. (2021). Rethinking perturbations in encoder-decoders for fast training. arXiv preprint arXiv:2104.01853.

[5] Ott M， Edunov S， Baevski A , et al. fairseq: A Fast, Extensible Toolkit for Sequence Modeling[C]// Proceedings of the 2019 Conference of the North. 2019.

[6] Sennrich R， Haddow B， Birch A . Neural Machine Translation of Rare Words with Subword Units[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

[7] Kingma D， Ba J . Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[8] Szegedy C， Vanhoucke V， Ioffe S , et al. Rethinking the Inception Architecture for Computer Vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016:2818-2826.

[9] Post M . A Call for Clarity in Reporting BLEU Scores[C]// Proceedings of the Third Conference on Machine Translation: Research Papers. 2018.

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

ICT's Submissions for CCMT 2022 Mongolian-Chinese Translation Task

Qingkai Fang[12†], Zhengrui Ma[12†], Shangtong Gui[12†], Zhuocheng Zhang[12],

Xuanfu Wu[12], Langlin Huang[12], Min Zhang[3], Yang Feng[12*]

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190; 2. University of Chinese Academy of Sciences, Beijing 100049; 3. Harbin Institute of Technology, Shenzhen, 518055)

Abstract: This paper describes the technical implementation of ICT's submissions for CCMT 2022 Mongolian-Chinese translation task. We first build a strong baseline system with deep Transformer architecture. To enhance the training data, we generate pseudo translations for monolingual Chinese corpus by back translation with dropout. The real parallel corpus and pseudo parallel corpus are combined together for training. To overcome the over-fitting and under-fitting problems, we use prediction difference regularization against perturbation to regularize the model. Finally, we ensemble several models during decoding. Experimental results show that our method achieves significant improvements over the strong baseline system.

Keywords: Neural Machine Translation; Back Translation; Regularization