中科院计算所 CCMT 2022 低资源神经翻译技术报告

张倬诚[12‡]，伍烜甫[12‡]，黄浪林[12‡]，房庆凯[12]，马铮睿[12]，桂尚彤[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与第十八届全国机器翻译大会的总体情况和技术细节。在本次评测中我们参加了泰语到汉语和汉语到泰语低资源翻译两个任务。针对两个任务的不同特点，本文给出了我们获取并处理语料、模型架构以及针对性的训练策略等方面的技术细节。实验表明，本次评测我们采用的策略和方法有效地提高了低资源情况下的翻译质量。

关键词：神经机器翻译；反向翻译；正则化；预训练模型；低资源机器翻译

中图分类号：TP391.2　　**文献标志码：** A

1 引言

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）低资源翻译任务的总体情况。该任务分为汉语到泰语方向（简称：CTH）和泰语到汉语方向（简称：THC）。

本次评测中，我们使用 Transformer 作为我们模型的基础架构，在使用 CCMT 提供的训练集之外，也在收集了部分公开的单语、双语语料。在语料处理过程中，我们使用了反向翻译、正向翻译等技术对数据集进行了扩充。在模型训练时，我们使用了预训练模型，且应用了计划采样、预测差异正则化、词表裁剪等技术来提高模型性能表现。最后，我们也利用了检查点平均以及模型集成方法来进一步提高模型的性能。在实验中，我们分析了不同技术对模型带来的性能提升，实验结果表明，本文应用技术均能有效提升模型性能。

2 数据的获取及处理

本节主要阐述评测中数据的获取及处理。由于本次评测任务中 CTH 及 THC 方向均未限制外部数据的使用，我们在使用 CCMT 官方提供数据的同时，也自行收集了部分公开的单语或双语语料。在此基础上，我们参考以往评测中的数据处理方法并结合此次收集数据的特点，对数据进行了预处理和清洗，得到了我们用于训练模型的语料。

2.1 数据获取

针对本次评测的特点，我们在获取语料时着重收集了泰语到汉语的通用平行语料、汉语和泰语的单语通用领域及新闻领域语料、泰语到英语的平行语料、英语到汉语的平行语料。具体地，我们在使用了来自 CCMT 官方提供的数据集之外，也收集了来自 OPUS[1]、WMT[2]等公开语料。此外，针对新闻领域语料较难获取的特点，我们额外收集了少量汉语及泰语的单语新闻语料。

[1] https://opus.nlpl.eu/

[2] https://www.statmt.org/wmt17/

基金项目：政府间国际科技创新合作重点专项（2017YFE0192900）

* **通信作者：**fengyang@ict.ac.cn

‡ 相同贡献

2.2 数据清洗及预处理

本次评测中，语料的清洗及预处理我们遵循了如下流程：

1. 规范化标点符号
2. 将句子中存在的繁体字转为简体（仅针对汉语）
3. 去除重复语料并去除空句子
4. 分词
5. 去除语料中含长度超过 20 的词的句子
6. 去除语料中长度小于 a 和大于 b 的句子
7. 将句子中存在的繁体标点符号转换为简体（仅针对汉语）
8. 删除语种识别与句子实际语种不符的句子
9. 删除包含特殊字符的句子
10. 删除困惑度过高的句子（仅针对汉语）

在实际实验中，我们分别采用了 sacremoses[3]、jieba[4]、polyglot[5]完成英语、汉语和泰语的分词，采用 pycld3[6]进行语种识别，采用 OpenCC[7]进行汉字的繁简转换，采用 GPT-2[15]对文本进行困惑度评价，采用 Moses[8]对标点符号进行规范化。

2.3 伪语料合成

本次评测中，我们使用了正向翻译、反向翻译及枢轴翻译技术来对平行语料进行扩增。其中正向翻译和反向翻译可以利用汉语及泰语的单语语料来生成汉语到泰语的伪平行语料，是各类评测中常见的数据增强方法；而枢轴翻译则是利用一个资源较为丰富的中间语言，将这个中间语言的数据集翻译到目标语言来获取质量更高的伪平行语料。

相较于平行语料，单语语料规模较大。基于以往研究，利用这些单语语料进行反向翻译可以对平行语料进行数据增强，进一步提高模型性能。由于汉语与泰语单语语料规模差异较大，直接进行反向翻译会导致以汉语为真实文本的语料规模远大于泰语为真实文本的语料。因此我们对在进行反向翻译前，从汉语和泰语中各筛选了约 1000 万条质量较高的句子进行反向翻译。在进行反向翻译时，我们首先利用现有的通用语料对 mBART[4]进行微调，得到了泰汉方向和汉泰方向两个模型，随后利用这两个模型对筛选出的句子进行翻译，再经过筛选后得到最终的反向翻译伪平行语料。

相较于汉泰方向，英泰方向以及英汉方向语料资源丰富，训练出的模型翻译性能通常比直接使用汉泰语料训练出的模型更好。因此我们利用现有的汉泰及英汉语料分别训练了英泰和英中两个模型。基于以上两个模型，我们对现有的泰汉语料及英汉语料做了以下处理：

1. 将汉泰语料中的泰语通过英泰模型翻译得到伪汉泰语料
2. 将英汉语料中的英文通过英中模型翻译得到伪汉泰语料

在进行以上两步处理后，我们将语料合并并利用生成时的困惑度进行进行了筛选，得到了最终的枢轴翻译伪平行语料。

2.4 领域筛选

本次评测中，我们收集了大量通用领域的语料，为了更好地使模型能够学习领域内知识，我们从大规模的通用领域语料中筛选出了少量的领域内数据。由于缺少文本的领域标签，我们无法直接训练一个分类器，因此我们通过领域内语言模型对语料进行领域筛选。根据文献[1]，若语料出现在领域内的对数概率与在整个数据集中出现的对数概率比值越大，则语料越有可能来自相应的领域。据

[3] https://github.com/alvations/sacremoses
[4] https://github.com/fxsjy/jieba
[5] https://github.com/aboSamoor/polyglot
[6] https://github.com/bsolomon1124/pycld3
[7] https://github.com/BYVoid/OpenCC
[8] https://github.com/moses-smt/mosesdecoder

此，在已知领域数据D_I的情况下，我们可以按照如下表达式筛选出语料s：

$$\arg max_s[\log P(s|D_I) - \log P(s|D)] \quad (1)$$

具体实现时，我们先利用领域内数据预训练一个语言模型，并利用该语言模型完成领域的筛选工作。

2.5 总结

经过上述语料收集及预处理过程，我们获得的全部语料类别及其规模如表 1 所示。

表 1 语料类型及规模

Tab.1 Type and Scale of Collected Corpus

语料类别	简称	语料规模
汉泰小说领域平行语料	CTH_{novel}	186K
汉泰通用领域平行语料	$CTH_{general}$	945K
汉泰枢轴翻译通用领域伪平行语料	$CTH_{general}^{pivot}$	18.4M
汉泰枢轴翻译新闻领域伪平行语料	CTH_{news}^{pivot}	2.0M
汉泰反向翻译通用领域伪平行语料	$CTH_{general}^{back}$	20.8M
汉泰反向翻译新闻领域伪平行语料	CTH_{news}^{back}	6.6M
汉英通用领域平行语料	$CE_{general}$	80.9M
英泰通用领域平行语料	$ETH_{general}$	5.1M
汉语新闻领域单语语料	C_{news}	355K
泰语新闻领域单语语料	TH_{news}	215K

3 模型结构

3.1 Transformer

谷歌 2017 年提出的神经机器翻译模型 Transformer[2]显著超过基于 RNN 方法的性能，至今仍是最主流的机器翻译模型。Transformer 模型由编码器和解码器构成，两个部分都采用注意力机制代替循环神经网络，缓解了长距离依赖问题，降低了训练时长。在本次评测中，我们采用了 12 层编码器+12 层解码器，隐向量维度 1024 的 Transformer-Large 模型。受限于硬件条件，我们无法选择更大规模的模型结构。

3.2 mBART

随着机器翻译及预训练技术的发展，大量多语言预训练语言模型被提出，如 T5[13]、mBART[4]、以及 M2M-100[11]。在本次评测中，我们选择预训练语言模型的原则如下：1. 能够同时支持泰语和中文；2. 提供 Transformer-Large 规模的检查点文件；3. 在非英语为中心的翻译任务中表现较好。基于以上原则，我们选择了 mBART50 作为预训练模型对检查点进行了初始化。

BART[3]是加入噪声的“编码器-解码器”结构的预训练语言模型，mBART[4]是同时训练了多种语言的预训练语言模型，它们都采用了 Transformer 模型结构进行训练，任务目标是把包含噪音的某种

语言句子还原。相对于 BART，mBART 的训练过程略有不同。首先，mBART 在训练过程中使用了多种语言的数据集；其次，为了使得模型能够区分不同语言的数据，在编码器端和解码器端额外输入了表示语种的语言标签，如<en>。实验结果[5]表明，利用多语言降噪任务训练得到的 mBART 模型可以有效学习到丰富的语言知识，在低资源语言方向上进行微调后翻译质量显著好于未采用 mBART 初始化的模型。

受限于预训练任务，mBART 仅能够支持 25 种语言。在此基础上，Meta 公司（原 Facebook）对 mBART 模型进行了扩展，训练了 mBART50[6]多对多模型。mBART50 在 mBART 的基础上进行了语种扩展，并引入平行语料进行训练，将预训练的 mBART 训练成一个可以完成多对多翻译的翻译模型。实验结果[6]表明，mBART50 能够有效地将单语预训练任务种学习到的知识迁移到翻译任务中。

3.3 词表裁剪

由于 mBART50 为了支持 50 种语言的文字，该模型的词表大小超过 25 万。但在本次评测中我们只会使用其中一小部分词语，冗余的词表一方面导致显存需求量激增，拖慢训练速度，另一方面在标签平滑过程中影响参数更新，可能对模型性能产生负面影响，因此我们对词表进行了裁剪。

具体地，我们检索了所有单语和双语语料中的泰语和汉语，仅保留在语料中出现过的词表单词。利用词表裁剪技术，我们成功使训练最大 batch size 扩大到了原来的 4 倍。

4 训练策略

4.1 总览

在本次评测中，我们利用前面所收集到的数据，对 mBART50 模型进行反复的训练和微调，以提升模型的性能。而在实验中，由于官方训练集和测试集的数据较少，我们采用了反向翻译和序列蒸馏的方法对数据进行了增强；同时为了改善正则化和缓解曝光偏差，我们分别在模型的训练中引入了预测差异正则化方法和计划采样。最终，我们通过模型集成的方式得到最终的模型并生成译文

本次评测的整体训练策略如图 1 所示，而根据 THC 和 CTH 方向领域的不同，我们在具体实验时又做了适应性的调整，从而使得模型更加贴近最终测试集所对应的领域。

4.2 泰汉方向

在训练泰汉方向时，我们将整个训练流程分为了三个阶段。第一阶段，我们在较大规模且通用领域的语料上进行训练，目的是希望模型学习到更丰富的语言学信息；随后我们在质量较高但规模较小的语料上进行第二阶段的训练，这一阶段主要是帮助模型纠正第一阶段从有噪声语料中学习到的错误知识；最后我们选择领域相关语料进行第三阶段的训练，提高目标领域的模型性能。

在第一阶段，我们选择汉泰枢轴翻译通用领域伪平行语料$\mathrm{CTH}_{\mathrm{general}}^{\mathrm{pivot}}$和汉泰反向翻译通用领域伪平行语料$\mathrm{CTH}_{\mathrm{general}}^{\mathrm{back}}$来训练模型，这两个数据集规模较大，涵盖领域丰富，有助于模型学习到更丰富的语言信息，为接下来的微调建立一个更好的基础。在第二阶段，我们利用汉泰通用领域平行语料$\mathrm{CTH}_{\mathrm{general}}$训练了泰语到汉语和泰语到汉语两个方向的模型，并利用这两个模型对汉泰通用领域平行语料$\mathrm{CTH}_{\mathrm{general}}$进行了数据增强。随后我们在增强的数据上进一步微调第一阶段训练得到的模型。在第三阶段，我们首先用汉泰小说领域平行语料训练了进行了泰语到汉语和泰语到汉语两个方向的模型，随后利用这两个模型对汉泰小说领域平行语料进行了数据增强，最后我们利用这增强后的汉泰小说领域平行语料$\mathrm{CTH}_{\mathrm{novel}}$继续对上一步训练得到模型进行进一步微调，得到了最终用于泰汉方向的模型。

最后，我们使用了模型集成技术对我们的模型进行了集成。

4.3 汉泰方向

与泰汉方向的训练过程相似，我们也将训练分为三个阶段。第一阶段与泰汉方向相同，使用大规模语料训练；第二阶段则使用从大规模语料中抽取的有噪声新闻领域训练集进行训练；第三阶段

则使用高质量的新闻领域单语语料进行迭代式正反向翻译，进一步提高模型在新闻领域上的表现。

具体地，我们在第二阶段使用了汉泰枢轴翻译新闻领域伪平行语料CTH_{news}^{pivot}和汉泰反向翻译新闻领域伪平行语料CTH_{news}^{back}进行训练，该语料规模较大，但依然存在大量噪声，无法被直接用于训练最终的模型。因此在第三阶段，我们选择了质量较高的汉语新闻领域单语语料C_{news}和泰语新闻领域单语语料TH_{news}进行迭代式正反向翻译来进一步提高模型的性能表现。

需要注意的是，由于我们未能获取到汉泰方向高质量的新闻领域数据，使用的数据均为通过正反向翻译或枢轴翻译得到的伪语料进行训练，这在一定程度上提高了我们训练模型的难度。

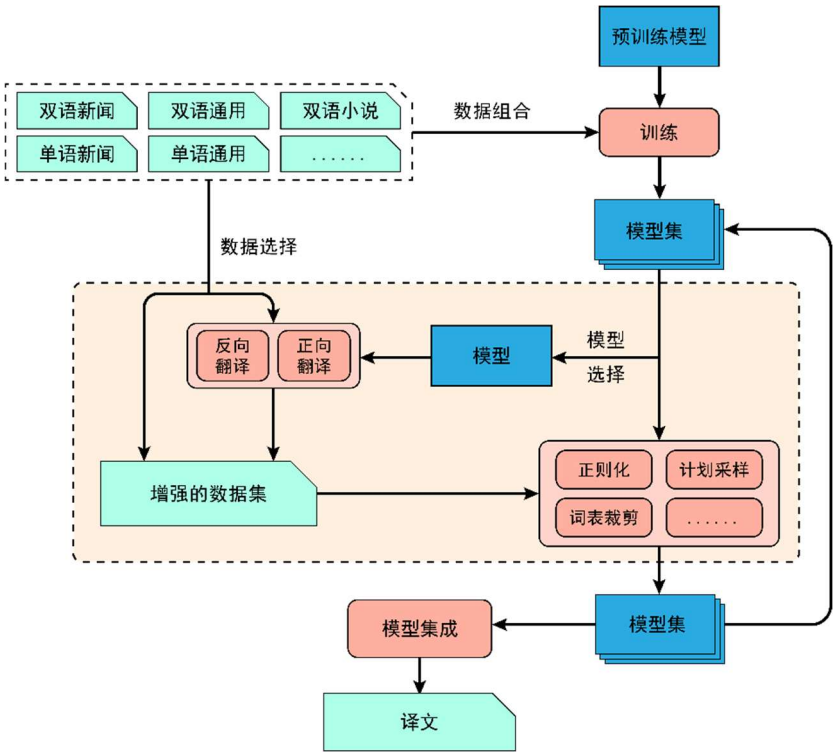

图 1 训练策略

Fig 1 Training Strategy

4.3 训练技术

在训练过程中，我们应用了以下几类训练技术来提高模型性能。

正反向翻译[7]：正反向翻译是机器翻译中常用的数据增强技术，通过训练正向及反向翻译模型，可以将源端及目标端单语数据翻译获得合成的平行语料。此外，我们也利用正反向翻译技术对现有平行语料进行了扩充，从而避免模型因数据量过小而陷入过拟合。

预测差异正则化[8]：在训练机器翻译模型时，由于模型容量及数据规模的不匹配，很容易出现如过拟合或欠拟合的现象。本文利用预测差异正则化(Prediction Difference Regularization)技术来缓解训练过程中产生的过拟合及欠拟合问题。该技术通过在训练过程中对训练数据动态添加扰动噪声，随后最小化有噪声数据和无噪声数据情况下模型的输出，从而使模型对数据中存在的微小差异鲁棒性更强。

计划采样[9, 10]：机器翻译中普遍存在着曝光偏差，即训练时模型采用的是参考译文的前 n-1 个词，而解码时采用的实际输出译文的 n-1 个词，两者分布存在着一定偏差。而 schedule sampling 则是为了解决这一问题的常用技术。

参数冻结[14]：在使用 Transformer 进行迁移学习时，我们通常使用模型的全部参数进行微调。但近年来的研究[14]表明，在进行领域迁移时，仅微调模型的部分参数就可以获得和微调全部模型相似的下游任务表现，同时减少模型的灾难性遗忘。具体的，我们冻结了模型的自注意力模块以及前馈神经网络部分，仅微调词嵌入向量和交叉注意力模块。

5 解码策略及后处理

5.1 模型集成

我们采用 fairseq 自带的模型集成方法对多个模型进行了集成，该方法在每个时间步用所有选取的模型分别预测输出的概率分布，再将它们按照指数方式进行平均得到最终的概率分布。

5.2 去未知词

由于 mBART50 预训练模型的词表中缺少部分汉语罕见字或异体字，我们的模型在 THC 方向进行解码生成时会出现未知词（UNK）。针对存在未知词的句子，我们利用在小说领域训练集上训练的前后向的三元文法语言模型预测未知词指代的词语，语言模型的解码过程会排除已经出现在 mBART50 词表中的词语。

6 实验

6.1 实验设置

在实验中，我们采用的硬件环境为：2 颗 Intel(R) Xeon(R) Silver 4214R 处理器，8 块 NVIDIA GeForce RTX 3090 显卡，256G 内存；软件环境为：Ubuntu 18.04 操作系统，CUDA 11.3，Python 3.9，Pytorch 1.11。

在实验中，我们采用了 Meta 公司（原 Facebook）开发的 fairseq 框架[9]。训练时，每个批次约 3 万个字符。优化器为 Adam，学习率为 5e-4，并保存最后 5 个检查点用于检查点平均。此外我们采用了 warmup 技术，并设置 warmup 步数为 4000。我们根据数据规模将实验参数分为 2 类：高资源（数据量高于一千万）、低资源（数据量低于一千万）。在高资源的情形下，我们设置 dropout=0.1，每个模型训练 10 万步；在低资源情形下，我们设置 dropout=0.2，每个模型训练 5 万步。翻译结果使用 fairseq 集成的 BLEU 进行评价。

6.2 测试集构建

由于本次评测并未提供测试集和开发集，因此在泰语到汉语的方向上我们从训练集中随机抽取了 426 条数据作为测试集，243 条作为验证集对模型进行评估；在汉语到泰语的方向上我们从 Flores[10] 多语言测试集和伪语料中选取了新闻部分作为测试集，其数据规模为 689 条。此外，为了衡量模型的泛化性能，我们也在部分实验中采用了从汉泰通用领域平行语料中抽取的测试集，规模为 3133 条；以及完整的 Flores 测试集，共计 1917 条。综上，我们所使用的测试集如表 2 所示。

表 2 测试集类别及规模

Tab.2 Type and Scale of test set

测试集类别	测试集规模
汉泰通用领域测试集	3133
汉泰 Flores 测试集	1917
汉泰小说领域测试集	426
汉泰新闻领域测试集	689

[9] https://github.com/facebookresearch/fairseq

[10] https://github.com/facebookresearch/fairseq/tree/main/examples/flores101

为了行文方便，我们在后文中将汉泰通用领域测试集和汉泰 Flores 测试集简称为：通用测试集、Flores 测试集，而汉泰小说领域测试集和汉泰新闻领域测试集则称之为领域测试集。

6.3 实验结果

在实验中，我们用 mBART50 作为预训练模型对模型进行了初始化，其核心原因是泰语与汉语均属于 mBART50 预训练中支持的语种，用 mBART 进行初始化可以给模型提供充分的先验知识。在训练泰语到汉语的实验中，我们也与未使用 mBART50 进行初始化的模型进行了对比，训练集为 $\mathrm{CTH}_{\mathrm{general}}$。结果如表 3 所示，使用 mBART50 初始化可以有效提升模型性能。

表 3 采用预训练模型的实验结果

Tab.3 The Results of using mBART

方向	模型设置	使用 mBART 初始化	领域测试集
THC	Transformer-base	否	54.64
	Transformer-large	否	56.82
	Transformer-large	是	**58.50**

在本次评测中，我们收集了一定规模的额外语料，也利用数据增强技术对原始语料及外部语料进行了增强。在使用了外部语料的情况下，模型性能有了进一步的增长。如表 4 所示，相对于领域特征较为明显的训练集，我们构建的外部语料可以大幅提高模型的泛化性能。

表 4 采用预训练模型的实验结果

Tab.4 The Results of using mBART

方向	训练集	通用测试集	领域测试集
THC	$\mathrm{CTH}_{\mathrm{novel}}$	26.37	**59.34**
	$\mathrm{CTH}_{\mathrm{general}}^{\mathrm{pivot}} + \mathrm{CTH}_{\mathrm{general}} + \mathrm{CTH}_{\mathrm{novel}}$	**58.41**	58.56

在进行模型训练时，我们采用了预测差异正则化、参数冻结、词表缩减等模型无关的训练技术，这些技术在不同方面对模型的性能产生了或多或少的影响，我们将其实验结果总结如表 5 所示。注意，这里我们仅在 THC 方向上进行了对照实验，训练集为$\mathrm{CTH}_{\mathrm{general}}$，但相关技术在 CTH 方向和 THC 方向上都有应用。

表 5 采用不同训练技术的实验结果

Tab.5 The Results of using different training techniques

方向	编号	训练技术	通用测试集	领域测试集
THC	1	基线	56.76	**58.50**
	2	1+部分参数冻结	55.76	57.30
	3	1+预测差异正则化	57.73	58.40
	4	3+词表缩减	**57.85**	58.45

可以看到，在使用了正则化以及词表缩减等技术后，模型在领域数据集上性能基本保持的情况下，通用领域测试集上有了明显的提升；而冻结部分参数方法则没有使用完整参数微调效果好。另外，由于官方提供的训练集领域特征明显且用词简单，使用基线模型就能够较好拟合，因此我们发现使用多种技术后领域测试集并没有进一步提高。

在训练泰语到汉语及汉语到泰语的过程中，我们都使用了多阶段微调的技术。使用该技术可以使得模型得以不断向目标领域靠近，从而获得较好的训练效果。此外，我们还使用了检查点平均以及模型集成方法来进一步提高模型性能。我们最终提交的系统结果如表 6 所示。

表 6 提交系统实验结果

Tab.6 The Results of submitted system

方向	系统名称	领域测试集
THC	thc-2022-ict-primary-a	61.73
CTH	cth-2022-ict-primary-a	59.60

7 总结

本次评测中我们尝试了从数据的获取及处理、模型的初始化、训练流程、训练技术、后处理技术等几个方面提升低资源方向机器翻译质量。此外，我们针对泰汉及汉泰两个方向设计了不同的策略来提升模型在不同场景下的性能表现。实验结果证明我们所采用的技术和策略均能够在数据有限的情况下提高模型的性能表现。

受限于时间和计算资源，我们在本次评测中未能尝试使用 M2M-100[11]等大型预训练模型，top-p 采样[12]等数据增强方法。在未来的工作和学习生活中我们也会进一步探索其它方法在低资源翻译领域的效果。

8 致谢

本次评测中课题组的各位同学和老师都给予了我们极大的帮助，通过此次评测，我们从他们身上学习到了大量的经验，也对机器翻译技术有了更深的理解。在此对他们致以诚挚的感谢。

参考文献：

[1] Moore R C, Lewis W. Intelligent selection of language model training data[C]//Proceedings of the ACL 2010 conference short papers. 2010: 220-224.

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

[3] Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 7871-7880.

[4] Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[5] Wang W, Jiao W, Hao Y, et al. Understanding and Improving Sequence-to-Sequence Pretraining for Neural Machine Translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 2591-2600.

[6] Tang Y , Tran C , Li X , et al. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning[J]. 2020.

[7] Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 86-96.

[8] Guo D, Ma Z, Zhang M, et al. Prediction Difference Regularization against Perturbation for Neural Machine Translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 7665-7675.

[9] Liu Y, Meng F, Chen Y, et al. Confidence-Aware Scheduled Sampling for Neural Machine Translation[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 2327-2337.

[10] Liu Y, Meng F, Chen Y, et al. Scheduled Sampling Based on Decoding Steps for Neural Machine Translation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 3285-3296.

[11] Fan A, Bhosale S, Schwenk H, et al. Beyond English-Centric Multilingual Machine Translation[J]. Journal of Machine Learning Research, 2021, 22: 1-48.

[12] Holtzman A, Buys J, Du L, et al. The Curious Case of Neural Text Degeneration[C]//International Conference on Learning Representations. 2019.

[13] Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. Journal of Machine Learning Research, 2020, 21: 1-67.

[14] Gheini M, Ren X, May J. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 1754-1765.

[15] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

ICT's Submissions for CCMT2022 Low Resource Neural Machine Translation Task

Zhuocheng Zhang[12‡], Xuanfu Wu[12‡], Langlin Huang[12‡], Qingkai Fang[12],

Zhengrui Ma[12], Shangtong Gui[12], Min Zhang[3], Yang Feng[12]

(1. Institute of Computing Technology, Chinese Academy of Science, Beijing, 100190, China; 2. University of Chinese Academy of Science, Beijing 100094; 3. Harbin Institute of Technology, Shenzhen, 518055)

Abstract: This paper describes the technical details of our participation in the 18th China Conference on Machine Translation. In this evaluation, we participated in two tasks of low-resource translation, which are from Thai to Chinese and from Chinese to Thai. According to the different characteristics of these tasks, this paper gives the technical details of our gathering and processing of corpus, model architecture, and training strategies. Experiments show that the strategies and methods we adopt in this evaluation can effectively improve the translation quality in low-resource situations.

Keywords: Neural Machine Translation, Back Translation, Regularization, Pretrain, Low Resource Machine Translation

‡ Equal Contribution