华为文本机器翻译实验室 CCMT 2021 双语翻译评测技术报告

吴章淋，李宗耀，於正哲，商恒超，郭嘉鑫，魏代猛，王明涵，雷立志，张敏，杨浩[1]*，秦璎

（华为文本机器翻译实验室，北京 100038）

摘要：本文详细介绍了华为文本机器翻译实验室（HWTSC）参加第十七届全国机器翻译大会机器翻译评测（CCMT 2021）的总体情况和采用的技术细节。在本次评测的翻译任务中，HWTSC 共参加了两个任务，分别是汉英新闻领域机器翻译和英汉新闻领域机器翻译任务。报告将主要阐述本次参评系统采用的模型框架、数据预处理方法以及模型增强策略。报告最后给出了在不同策略下评测系统在评测数据上的性能表现，并进行了对比和分析。

关键词：机器翻译；CCMT 2021；神经网络

中图分类号：TP391　　文献标志码：A

1 引言

本文详细介绍了华为文本机器翻译实验室（HWTSC）参加第十七届全国机器翻译大会机器翻译评测（CCMT 2021）翻译评测任务的总体情况。在CCMT 2021的双语翻译任务类型下，HWTSC只参加了汉英新闻领域机器翻译和英汉新闻领域机器翻译任务。

本次评测采用谷歌 Transformer[1]神经网络机器翻译架构。在汉英新闻领域机器翻译和英汉新闻领域机器翻译任务上，使用了 WMT 2021 和 CCMT 2021 的汉英双语和单语。在数据预处理方面，针对评测方发布的数据，一方面，采取多种不同语料过滤方法减少语料噪声以提高训练语料的质量，另一方面，训练领域分类器以选取更接近新闻领域的语料。在训练策略方面，使用了正向翻译和反向翻译的方法来构建伪平行语料，以扩大训练集的规模，此外，还使用了模型微调和集成解码策略对模型进行了进一步的优化。在实验中，对比了模型在不同的数据增强策略下的表现，并对实验结果进行了分析。

2 系统介绍

本次评测使用的基线系统是基于自注意力机制的 Transformer[1]模型，模型采用完全自注意力机制，能够在实现算法并行性、加快模型训练速度的同时，提高翻译质量。本次评测使用了三种 Transformer 模型架构：

1) deep 25-6[2,3]模型：在 Transformer-base 模型架构的基础上，编码器的层数调整为 25 层，并前置层归一化。
2) deep 35-6[2,3]模型：在 Transformer-base 模型架构的基础上，编码器的层数调整为 35 层，并前置层归一化。
3) deep 25-6 large[4,5]模型：词向量维度为 768，隐层状态维度为 3072，编码器为 25 层，解码器为 6 层，多头自注意力机制使用 16 个头，并前置层归一化。

3 方法介绍

本次评测中，HWTSC 共参加了汉英新闻领域机器翻译和英汉新闻领域机器翻译任务。两个评测任务使用的方法是一致的，在采用基于自注意力机制的 Transformer 模型的基础上，

* 通信作者：yanghao30@huawei.com

使用了数据选择、数据增强、模型微调、集成解码方法。

3.1 数据选择

数据选择旨在挑选更接近本次评测领域的语料，本次评测使用新闻单语和非新闻单语，分别训练基于fasttext[6]的汉语二分类模型和英语二分类模型。数据选择包括双语和单语两个方面，在选择单语时，用其对应语言的二分类模型打分排序，选取在新闻领域得分高的语料。在选择双语时，源测和目标测分别用其对应语言的二分类模型进行打分，然后取其平均得分作为最终得分，以此为依据，选取在新闻领域得分高的语料。

3.2 数据增强

与双语数据相比，单语数据的规模是庞大的，单语数据增强对模型的质量提升至关重要。本次评测对比了几种常用的单语增强策略，包括基于beam search加标签的反向翻译[7]、基于top-k随机采样的反向翻译[8]、基于beam search的正向翻译和反向翻译的组合[9]以及基于beam search的正向翻译和基于top-k随机采样的反向翻译的组合。在双语和单语数据规模悬殊的场景下，基于beam search的正向翻译和基于top-k随机采样的反向翻译的组合可以取得最优的效果。本次评测最终采用了基于beam search的正向翻译和基于top-k随机采样的反向翻译的组合策略。

3.3 模型微调

为了使模型在测试集上表现良好，在同领域的小数据集上对模型进行微调是有必要的。直接使用本次评测提供的开发集对模型微调是一种常用的策略，为了衡量模型在开发集上的表现，本次评测使用多个模型对测试集的源文进行集成解码，然后对目标测加噪处理[10]，构造训练数据，用于微调，也可以取得不错的效果。

3.4 集成解码

本次评测随机了2份数据，并使用上文提到的三种模型架构训练了6个模型，选取在开发集上bleu较高的4个模型进行了集成解码[11]。

4 实验

本评测使用的系统为 fairseq[12]开源工具。主要参数设置如下，每个模型使用 8 块 GPU 进行训练，每个 batch 大小为 2048，参数更新频率设置为 32，学习率为 5e-4，warmup 步数为 4000。deep 25-6 模型和 deep 35-6 模型的词向量维度为 512，隐层状态维度为 2048，解码器为 6 层，多头自注意力机制使用 8 个头，编码器的层数分别对应为 25 层和 35 层。deep 25-6 large 模型的词向量维度为 876，隐层状态维度为 3072，编码器层为 25 层，解码器为 6 层，多头自注意力机制使用 16 个头。本次评测采用了 dropout 机制，dropout 设为 0.1。训练语料先分词再用 bpe[13]切分，源语言及目标语言的词表共享设定为 32K，汉语分词采用 jieba[14]，英语分词采用 Moses。在推理阶段，本评测采用 marian[15]工具进行解码，beam-size 设置为 10， zh2en 的长度惩罚设置为 1.2，而 en2zh 的长度惩罚设置为 0.8。此外，在基于 top-k 随机采样制造伪语料时，使用的是 fairseq[12]进行解码，解码参数设置为 beam=1，sampling=True，sampling_topk=10。

4. 1 数据处理

本次评测的汉英新闻领域机器翻译和英汉新闻领域机器翻译任务使用了 CCMT 2021 和 WMT'2021 提供的平行语料和单语数据，表 1 为使用的详细数据情况。

4.1.1 语料预处理

汉英双语主要的数据处理过程为：

1) 去除重复行
2) Moses[16]标点归一化
3) xml 转义符还原
4) 空格修正
5) 去除非 utf-8 字符、html 标签、unicode 字符以及不可见字符
6) 汉语繁体转换为简体
7) langid[17]过滤
8) 全角符号转换为半角符号
9) 标点最大占比超过 0.3 过滤
10) 括号和引号不匹配过滤
11) 字符和单词比例大于 12 或小于 1.5 过滤
12) 句子单词数大于 120 过滤
13) 句中存在重复片段过滤
14) fast-align[18]过滤

最后，再用 fasttext 工具[6]进行领域数据选择，最终选取了 16.5M 双语。

4.1.2 单语数据

汉英单语主要的数据处理过程为：

1) langid[17]语言过滤
2) 句子单词数大于 120 过滤
3) Kenlm[19]过滤长度归一化后得分较低的句子

通过数据处理和领域数据选择，本次评测最终选取了汉语和英语单语各 150M。

表 1 本次评测使用的数据情况

Tab.1 The data used in this assessment

数据类型	双语			汉语单语		英语单语
数据来源	CCMT Corpus & News Commentary v16	Wiki Titles v3 & WikiMatrix	ParaCrawl v7.1 & UN Parallel Corpus V1.0	News crawl & News Commentary	Common Crawl	News crawl
初始数据量	7.3M	3.4M	27M	10.4M	1672M	260M
最终数据量	16.5M			150M		150M

4. 2 实验结果

4.2.1 主要结果

表 2 和表 3 分别为本次评测提交英汉和汉英的机器翻译模型在开发集上的评测结果。测评指标采用的是 sacrebleu[20]，ccmt2019-news 1ref 表示的是用参考译文 1 计算的 BLEU 结果，ccmt2019-news 4ref 表示的是用 4 个参考译文计算的 BLEU 结果， baseline 是指由 16.5M 双

语训练出来的 deep 25-6 模型，FTST 是指由 16.5M 双语、150M 汉语伪语料以及 150M 英语伪语料训练出来的模型，伪语料的正向翻译使用基于 beam search 的方法，而反向翻译使用基于 top-k 随机采样，In-domain_FTST 是指用测试集微调得到新的模型，然后制造伪语料，训练第二轮 FTST deep 25-6 模型，finetune 是指在上一步的基础上用 CCMT 2019-2021 的测试集源文造伪语料，并对目标测进行加噪处理，最后用于微调，在汉英任务上，微调的训练语料还加入了 WMT 2017-2018 测试集，ensemble 是指 4 个模型集成解码。

从表 2 和表 3 可以看出知识迁移的迭代式数据增强、模型微调以及集成解码的策略在 CCMT 2021 汉英新闻领域机器翻译和英汉新闻领域机器翻译任务上的效果。

表 2 英汉翻译评测在开发集上的结果

Tab.2 The results of the English-Chinese translation evaluation on the development set

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
baseline	43.30	65.01
FTST	47.30	71.15
In-domain_FTST	47.76	71.61
finetune	48.22	72.29
ensemble	**48.28**	**72.41**

表 3 汉英翻译评测在开发集上的结果

Tab.3 The results of the Chinese- English translation evaluation on the development set

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
baseline	27.16	45.06
FTST	31.85	52.25
In-domain_FTST	35.83	58.54
finetune	36.40	59.21
ensemble	**36.59**	**59.50**

4.2.2 不同单语增强策略对模型质量的影响

本次评测还对同一数据集下不同单语增强策略对模型质量的影响进行了研究。表 4 和表 5 分别为英汉和汉英的机器翻译模型在开发集上的评测结果。模型架构均采用 deep 25-6，TAGBT 是指由 16.5M 双语和 150M 译文单语基于 beam search 制造的伪语料，并给伪语料源文加标签训练出来的模型，ST 是指由 16.5M 双语和 150M 译文单语基于 top-k 随机采样制造的伪语料训练出来的模型，FTBT 是指由 16.5M 双语、150M 汉语伪语料以及 150M 英语伪语料训练出来的模型，伪语料的正向翻译和反向翻译均使用基于 beam search 的解码策略，FTST 是指由 16.5M 双语、150M 汉语伪语料以及 150M 英语伪语料训练出来的模型，伪语料的正向翻译使用基于 beam search 的解码策略，而反向翻译使用基于 top-k 随机采样的解码策略。

从表 4 和表 5 的结果可以看出，在单语数据规模远超双语的场景下，FTST 策略可以取得很好的效果。

表 4 在不同单语数据增强策略下，英汉翻译评测在开发集上的结果

Tab.4 The results of the English-Chinese translation evaluation on the development set under Different Monolingual Enhancement Strategies

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
TAGBT	45.24	68.26
ST	46.26	69.22
FTBT	46.98	70.69
FTST	**47.30**	**71.15**

表 5 在不同单语数据增强策略下，汉英翻译评测在开发集上的结果

Tab.5 The results of the English-Chinese translation evaluation on the development set under Different Monolingual Enhancement Strategies

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
TAGBT	28.26	47.41
ST	28.76	47.10
FTBT	31.61	51.80
FTST	**31.85**	**52.25**

4.2.3 不同模型架构对模型质量的影响

此外，本评测还对不同模型架构在同一数据集下对模型质量的影响进行了研究。表 6 和表 7 分别为英汉和汉英的机器翻译模型在开发集上的评测结果。训练数据均采用 In-domain_FTST 语料，deep 25-6、deep 35-6 以及 deep 25-6 large 分别表示由对应模型架构在同一训练集下训练出来的模型。

表 6 在不同模型架构下，英汉翻译评测在开发集上的结果

Tab.6 The results of the English-Chinese translation evaluation on the development set under Different Model Frameworks

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
deep 25-6	**47.76**	**71.61**
deep 35-6	47.63	71.45
deep 25-6 large	47.69	71.46

表 7 在不同模型架构下，汉英翻译评测在开发集上的结果

Tab.7 The results of the English-Chinese translation evaluation on the development set under Different Model Frameworks

模型策略	ccmt2019-news 1ref	ccmt2019-news 4ref
deep 25-6	35.83	58.54
deep 35-6	35.77	58.26
deep 25-6 large	**35.98**	**58.63**

从表 6 和表 7 的结果可以看出，在该场景下，这三种不同架构的模型质量较为接近，测试集的 Bleu 值相差很小。

5 总结

本文详细介绍了华为文本机器翻译实验室（HWTSC）在CCMT 2021汉英新闻领域机器翻译和英汉新闻领域机器翻译任务上使用的主要技术和方法。总的来说，本次评测在模型上使用了基于自注意力机制的transformer的架构，在数据处理和选择方面，探索了多种语料过滤和选择方法，在训练策略方面，采用了知识迁移的迭代式数据增强、模型微调以及集成解码的策略，实验结果证明，这些方法能够有效提高翻译质量。由于时间有限，本次评测中还有一些方法没有尝试，在评测过程中也发现了一些问题和不足。在今后的研究中期望能够学习更多先进技术，对机器翻译研究有所贡献。

参考文献：

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[2] Wang Q, Li B, Liu J, et al. The niutrans machine translation system for wmt18[C]//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. 2018: 528-534.

[3] Li B , Li Y , Xu C , et al. The NiuTrans Machine Translation Systems for WMT19[C]// Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019.

[4] Wu L, Pan X, Lin Z, et al. The Volctrans Machine Translation System for WMT20[J]. arXiv preprint arXiv:2010.14806, 2020.

[5] Sun M, Jiang B, Xiong H, et al. Baidu neural machine translation systems for WMT19[C]//Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). 2019: 374-381.

[6] Joulin A, Grave E, Bojanowski P, et al. Fasttext. zip: Compressing text classification models[J]. arXiv preprint arXiv:1612.03651, 2016.

[7] Caswell I, Chelba C, Grangier D. Tagged back-translation[J]. arXiv preprint arXiv:1906.06442, 2019.

[8] Edunov S, Ott M, Auli M, et al. Understanding back-translation at scale[J]. arXiv preprint arXiv:1808.09381, 2018.

[9] Wu L, Wang Y, Xia Y, et al. Exploiting monolingual data at scale for neural machine translation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 4198-4207.

[10] Meng F, Yan J, Liu Y, et al. WeChat Neural Machine Translation Systems for WMT20[J]. arXiv preprint arXiv:2010.00247, 2020.

[11] Wang Y, Wu L, Xia Y, et al. Transductive ensemble learning for neural machine translation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6291-6298.

[12] Ott M, Edunov S, Baevski A, et al. fairseq: A fast, extensible toolkit for sequence modeling[J]. arXiv preprint arXiv:1904.01038, 2019.
[13] Provilkov I, Emelianenko D, Voita E. Bpe-dropout: Simple and effective subword regularization[J]. arXiv preprint arXiv:1910.13267, 2019.
[14] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.
[15] Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, et al. Marian: Fast neural machine translation in C++[J]. arXiv preprint arXiv:1804.00344, 2018.
[16] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation[C]//Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions. 2007: 177-180.
[17] Lui M, Baldwin T. langid. py: An off-the-shelf language identification tool[C]//Proceedings of the ACL 2012 system demonstrations. 2012: 25-30.
[18] Dyer C, Chahuneau V, Smith N A. A simple, fast, and effective reparameterization of ibm model 2[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013: 644-648.
[19] Heafield K. KenLM: Faster and smaller language model queries[C]//Proceedings of the sixth workshop on statistical machine translation. 2011: 187-197.
[20] Post M. A call for clarity in reporting BLEU scores[J]. arXiv preprint arXiv:1804.08771, 2018.

HWTSC's Submissions for CCMT 2021 Shared Translation Tasks

Zhanglin Wu, Zongyao Li, Zhengzhe Yu, Hengchao Shang, Jiaxin Guo, Daimeng Wei, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang [1]*, Ying Qin

(Text Machine Translation Lab, Huawei Beijing Research Center, Beijing 100038, China)

Abstract: This paper presents the systems developed and techniques adopted by Huawei Text Machine Translation Lab (HWTSC) for CCMT shared translation tasks: Chinese-English and English-Chinese news translation. In this paper, we explain our model framework, data preprocessing methods and model enhancement strategies. The performances of our systems under different strategies are also presented at the end of the paper, along with comparison and analysis.

Key Words: Machine Translation; CCMT 2021; Neural Networks