Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,weidaimeng,lizongyao,yuzhengzhe,lishaojun18,chenxiaoyu35, shanghengchao,guojiaxin1,xieyuhao2,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the submission of Huawei Translate Services Center (HW-TSC) to the WMT23 general machine translation (MT) shared task. We participate in Chinese↔English (zh↔en) language pair. We use deep Transformer architecture and obtain the best performance via a Transformer variant with a larger parameter size. We perform fine-grained pre-processing and filtering on the provided large-scale bilingual and monolingual datasets. The model enhancement strategies we used includes Regularized Dropout, Bidirectional Training, Data Diversification, Forward Translation, Back Translation, Alternated Training, Curriculum Learning and Transductive Ensemble Learning. Our submission obtain competitive results in the final evaluation.

1 Introduction

Machine translation (MT) (Brown et al., 1990) refers to the automatic translation of text from one language to another, while the WMT23 general MT shared task focuses on evaluation of general MT capabilities. Compared with the news shared task in previous years, the general MT shared task involves multiple domains. The testsets contain data in news, user generated (social), conversational, and ecommerce domains.

This paper presents the submission of HW-TSC to the WMT23 general MT shared task, in which we participate in zh↔en language pair. Our method is mainly based on previous works (Wei et al., 2022; Wu et al., 2022; Yang et al., 2021). We perform multi-step data cleansing on the provided dataset and only keep a high-quality subset for training. At the same time, several model enhancement strategies are tested in a pipeline, including Regularized Dropout (Wu et al., 2021), Bidirectional Training (Ding et al., 2021), Data Diversification (Nguyen et al., 2020), Forward Translation (Abdulmumin, 2021), Back Translation (Sennrich et al., 2016), Alternated Training (Jiao et al., 2021), Curriculum Learning (Zhang et al., 2019) and Transductive Ensemble Learning (Wang et al., 2020b).

Our system report includes four parts. Section 2 focuses on our data processing strategies while section 3 describes our training details. Section 4 explains our experiment settings and training processes and section 5 presents the results.

2 Data

2.1 Data Source

We obtain bilingual and monolingual data from ParaCrawl v9, News Commentary v18.1, Wiki Titles v3, UN Parallel Corpus V1.0, CCMT Corpus, WikiMatrix, News Crawl and Common Crawl data sources. The amount of data we used is shown in Table 1. It should be noted that in order to obtain better performance in the general domain, we mix the monolingual data from Common Crawl and News Crawl.

language pairs	bitext data	monolingual data
zh↔en	25M	en: 50M, zh: 50M

Table 1: Bilingual and monolingual used for training.

2.2 Data Pre-processing

Our data processing procedure is precisely the same as the previous year (Wei et al., 2021), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies. As we use the same data pre-processing strategy as the previous year, we will not go into details here.

2.3 Data Denoising

Since there may be some semantically dissimilar sentence pairs in bilingual data, we use LaBSE (Feng et al., 2022) to calculate the semantic similarity of each bilingual sentence pair, and exclude

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 170–174
December 6–7, 2023. ©2023 Association for Computational Linguistics

bilingual sentence pairs with a similarity score lower than 0.7 from our training corpus.

3 System Overview

3.1 Model

We continue using Transformer (Vaswani et al., 2017) as our neural machine translation (NMT) (Bahdanau et al., 2015) model architecture. As we did last year, we only use a 25-6 deep model architecture (Wang et al., 2019). The parameters of the model are the same as Transformer big. We just change the post-layer normalization to the pre-layer normalization, and set encoder layers to 25.

3.2 Regularized Dropout

Regularized Dropout (R-Drop)[1] (Wu et al., 2021) is a simple yet more effective alternative to regularize the training inconsistency induced by dropout (Srivastava et al., 2014). Concretely, in each mini-batch training, each data sample goes through the forward pass twice, and each pass is processed by a different sub model by randomly dropping out some hidden units. R-Drop forces the two distributions for the same data sample outputted by the two sub models to be consistent with each other, through minimizing the bidirectional Kullback-Leibler (KL) divergence (Van Erven and Harremos, 2014) between the two distributions. That is, R-Drop regularizes the outputs of two sub models randomly sampled from dropout for each data sample in training. In this way, the inconsistency between the training and inference stage can be alleviated.

3.3 Bidirectional Training

Many studies have shown that pre-training can transfer the knowledge and data distribution, hence improving the model generalization. Bidirectional training (BiT) (Ding et al., 2021) is a simple and effective pre-training method for NMT. Bidirectional training is divided into two stages: (1) bidirectionally updates model parameters, and (2) tune the model. To achieve bidirectional updating, we only need to reconstruct the training samples from "src→tgt" to "src→tgt & tgt→src" without any complicated model modifications. Notably, BiT does not require additional parameters or training steps and only uses parallel data.

3.4 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a data augmentation method to boost NMT performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset which the final NMT model is trained on. DD is applicable to all NMT models. It does not require extra monolingual data, nor does it add more parameters. To conserve training resources, we only use one forward model and one backward model to diversify the training data.

3.5 Forward Translation

Forward translation (FT) (Abdulmumin, 2021), also known as self-training, is one of the most commonly used data augmentation methods. FT has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. Generally, FT is performed in three steps: (1) randomly sample a subset from the large-scale source monolingual data; (2) use a “teacher” NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a “student” NMT model.

3.6 Back Translation

An effective method to improve NMT with target monolingual data is to augment the parallel training data with back translation (BT) (Sennrich et al., 2016; Wei et al., 2023). There are many works expand the understanding of BT and investigates a number of methods to generate synthetic source sentences. Edunov et al. (2018) find that back translations obtained via sampling or noised beam outputs are more effective than back translations generated by beam or greedy search in most scenarios. Caswell et al. (2019) show that the main role of such noised beam outputs is not to diversify the source side, but simply to tell the model that the given source is synthetic. Therefore, they propose a simpler alternative strategy: Tagged BT. This method uses an extra token to mark back translated source sentences, which generally outperforms noised BT (Edunov et al., 2018). For better joint use with FT, we use sampling back translation (ST) (Edunov et al., 2018).

[1] https://github.com/dropreg/R-Drop

3.7 Alternated Training

While synthetic bilingual data have demonstrated their effectiveness in NMT, adding more synthetic data often deteriorates translation performance since the synthetic data inevitably contains noise and erroneous translations. Alternated training (AT) (Jiao et al., 2021) introduce authentic data as guidance to prevent the training of NMT models from being disturbed by noisy synthetic data. AT describes the synthetic and authentic data as two types of different approximations for the distribution of infinite authentic data, and its basic idea is to alternate synthetic and authentic data iteratively during training until the model converges.

3.8 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) method should address two main questions: how to rank the training examples, and how to modify the sampling procedure based on this ranking. For ranking, we choose to estimate the difficulty of training samples according to their domain feature (Wang et al., 2020a). The calculation formula of domain feature is as follows, where θ_{in} represents an in-domain NMT model, and θ_{out} represents a out-of-domain NMT model. One thing to note is that we treat domains including news, user-generated (social), conversational, and e-commerce domains as in-domain, and others as out-of-domain. Specifically, we use the WMT22 test set to fine-tune a baseline model, and then use the baseline model and the fine-tuned model as the out-of-domain model and the in-domain model respectively.

$$q(x, y) = \frac{\log P(y|x; \theta_{in}) - \log P(y|x; \theta_{out})}{|y|} \tag{1}$$

For sampling, we adopt a probabilistic CL strategy that leverages the concept of CL in a nondeterministic fashion without discarding the original standard training practice, such as bucketing and mini-batching.

3.9 Transductive Ensemble Learning

Ensemble learning (Garmash and Monz, 2016), which aggregates multiple diverse models for inference, is a common practice to improve the performance of machine learning models. However, it has been observed that the conventional ensemble methods only bring marginal improvement for NMT when individual models are strong or there are a large number of individual models. Transductive Ensemble Learning (TEL) (Zhang et al., 2019) studies how to effectively aggregate multiple NMT models under the transductive setting where the source sentences of the test set are known. TEL uses all individual models to translate the source test set into the target language space and then finetune a strong model on the translated synthetic data, which significantly boosts strong individual models and benefits a lot from more individual models.

4 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training, then we use SacreBLEU (Post, 2018)[2] and wmt20-comet-da model (Rei et al., 2020) to measure system performances. The main parameters are as follows: each model is trained using 8 A100 GPUs, batch size is 6144, parameter update frequency is 2, and learning rate is 5e-4. The number of warmup steps is 4000, and model is saved every 1000 steps. The architecture we used is described in section 3.1. We adopt dropout, and the rate varies across different training phases. R-Drop is used in model training, and we set λ to 5.

5 Results

Regarding zh↔en, we use Regularized Dropout, Bidirectional Training, Data Diversification, Forward Translation, Back Translation, Alternated Training, Curriculum Learning, and Transductive Ensemble Learning. The evaluation results of en→zh and zh→en NMT system on WMT22 general test sets are shown in Tables 2.

	en→zh		zh→en	
	BLEU	COMET	BLEU	COMET
BiT R-Drop baseline	45.55	50.24	22.30	22.28
+ DD, FT & ST	49.54	59.69	25.67	33.44
+ AT	54.11	63.99	28.58	37.15
+ CL	56.36	68.90	30.58	44.62
+ TEL	**56.80**	**69.06**	**31.35**	**45.56**

Table 2: BLEU and COMET scores of en→zh and zh→en NMT system on WMT22 general test set.

We observe that DD, FT & ST can stably bring 3-4 BLEU and 1-9 COMET improvement; AT can bring 3-5 BLEU and 4 COMET improvement; and CL can bring 2 BLEU and 5-7 COMET improvement. In addition, TEL can further slightly improve BLEU and COMET scores. Our final en→zh

[2] https://github.com/mjpost/sacrebleu

System	chrF	BLEU	COMET
HW-TSC	**57.5**	**33.6**	**82.8**
ONLINE-B	**57.5**	33.5	82.7
Yishu	57.4	33.4	82.7
GPT4-5shot	53.1	26.8	81.6
Lan-BridgeMT	53.1	27.3	81.2
ONLINE-G	53.9	26.6	80.9
ONLINE-Y	52.3	25.0	80.6
ONLINE-A	53.4	28.3	80.3
ZengHuiMT	54.6	27.0	79.6
ONLINE-W	52.5	26.4	79.3
IOL_Research	52.4	27.2	79.2
ONLINE-M	49.7	23.5	77.7
NLLB_MBR_BLEU	45.8	19.8	76.8
ANVITA	47.1	21.8	76.6
NLLB_Greedy	46.1	20.5	76.4

Table 3: Scores for the WMT23 zh→en translation task: chrF, BLEU and COMET (Unbabel/wmt22-comet-da).

System	chrF	BLEU	COMET
ONLINE-B	52.9	57.5	**88.1**
Yishu	53.0	57.6	**88.1**
HW-TSC	**53.8**	**58.6**	87.3
GPT4-5shot	46.5	49.6	87.1
ONLINE-W	47.3	52.1	86.8
Lan-BridgeMT	46.8	50.2	86.6
ONLINE-Y	49.8	54.2	86.5
ONLINE-A	52.8	58.5	86.2
IOL_Research	51.9	56.9	85.3
ZengHuiMT	47.0	52.9	84.3
ONLINE-M	50.6	54.9	84.2
ONLINE-G	49.4	54.1	83.8
NLLB_Greedy	26.3	27.4	75.7
ANVITA	36.9	38.9	75.6
NLLB_MBR_BLEU	21.1	19.1	71.5

Table 4: Scores for the WMT23 en→zh translation task: chrF, BLEU, COMET (Unbabel/wmt22-comet-da).

and zh→en submissions achieve 56.80 and 31.35 BLEU, 69.06 and 45.56 COMET respectively.

6 Official Automatic Evaluation Results

In our final submission, we add post-processing for punctuation correction and entity preservation. WMT (Kocmi et al., 2023) present an automatic evaluation of the systems submitted to the general machine translation task, including the following three different automatic metrics: chrF, BLEU and COMET. We rank the systems according to COMET scores, and unconstrained systems are in a grey background in the tables.

7 Conclusion

This paper presents the submission of HW-TSC to the WMT23 general MT Task. We participate in zh↔en language pair and perform experiments with a series of pre-processing and training strategies. The effectiveness of each strategy is demonstrated. Our experiments show that our model training strategies are effective. Our submission finally achieve competitive results in the evaluation.

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. Alternated training with synthetic and authentic data for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020b. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.