# HW-TSC's Submissions to the WMT 2022 General Machine Translation Shared Task

**Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, Ying Qin,**
Huawei Translation Service Center, Beijing, China
{weidaimeng,raozhiqiang,wuzhanglin2,lishaojun18,luoyuanchang, xieyuhao2,chenxiaoyu35,shanghengchao,lizongyao,yuzhengzhe, yangjinlong7,mamiaomiao,leilizhi,yanghao30,qinying}@huawei.com

## Abstract

This paper presents the submissions of Huawei Translate Services Center (HW-TSC) to the WMT 2022 General Machine Translation Shared Task. We participate in 6 language pairs, including Zh↔En, Ru↔En, Uk↔En, Hr↔En, Uk↔Cs and Liv↔En. We use Transformer architecture and obtain the best performance via multiple variants with larger parameter sizes. We perform fine-grained pre-processing and filtering on the provided large-scale bilingual and monolingual datasets. For medium and high-resource languages, we mainly use data augmentation strategies, including Back Translation, Self Training, Ensemble Knowledge Distillation, Multilingual, etc. For low-resource languages such as Liv, we use pre-trained machine translation models, and then continue training with Regularization Dropout (R-Drop). The previous mentioned data augmentation methods are also used. Our submissions obtain competitive results in the final evaluation.

## 1 Introduction

This paper introduces our submissions to the WMT 2022 General Machine Translation Shared Task. We participate in 6 language pairs including Chinese/English (Zh↔En), Russian/English (Ru↔En), Ukrainian/English (Uk↔En), Croatian/English (En→Hr), Ukrainian/Czech(Uk↔Cs), and Livonian/English (Liv↔En). For Zh↔En translation, we use additional in-house in-domain data, so the final submission for this language pair is unconstrained. For Liv↔En translation, although we did not use additional data, we used M2M-100 (Fan et al., 2020) as the pretrained model, and the final submission is also unconstrained. All other languages pair participate in the constrained evaluation. Our method is mainly based on previous works (Wei et al., 2020, 2021; Yang et al., 2021) but with fine-grained data cleansing techniques and language-specific optimizations.

For each language pair, we perform multi-step data cleansing on the provided dataset and only keep a high-quality subset for training. At the same time, several strategies are tested in a pipeline, including Backward (Edunov et al., 2018) and Forward (Wu et al., 2019a) Translation, Multilingual Translation (Johnson et al., 2017), Iterative Joint Training (Zhang et al., 2018), R-Drop, Pretrained NMT model, Ensemble Knowledge Distillation (Freitag et al., 2017; Li et al., 2019), Fine-Tuning (Sun et al., 2019), Ensemble (Garmash and Monz, 2016), and Post-Processing.

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 data sources such as CCMT, UN, ParaCrawl, WikiMatrix, WikiTitles, News Commentary, Leipzig Corpora, News Crawl, and Common Crawl. 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.

### 2.2 Data Pre-processing

Our data processing procedure is basically the same as our method last year (Wei et al., 2021), including deduplication, XML content processing, langid (Joulin et al., 2016b,a) and fast-align (Dyer et al., 2013) filtering strategies, etc. As we use the same data pre-processing strategy as last year's, we will not go into details here.

### 2.3 Data Denoise

Regarding Hr↔En, the CCMatrix data is highly noisy, so more fine-grained data cleaning is nec-



*Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 403–410
Abu Dhabi, December 7–8, 2022. 

| language pairs | Raw bi data | Filter bi data | Used mono data |
|---|---|---|---|
| Zh/En | 39M | 37M | En: 150M (C&N), Zh: 150M (C) |
| Ru/En | 28M | 26M | En: 160M (C&N), Ru: 160M (C&N) |
| Hr/En | 69M | 55M | Hr: 22M (N) |
| Uk/En | 39M | 36M | En: 150M (C&N), Uk: 60M (N) |
| Cs/Uk | 8.4M | 8M | Cs: 60M (C&N), Uk: 60M (N) |
| Liv/En | 1.1k | 1.1k | Liv: 50K, En: 1M |

Table 1: Bilingual data sizes before and after filtering, and monolingual data used in the task. Regarding monolingual data, **N** means that the data comes from News Crawl; **C** means that the data comes from Common Crawl; and **C&N** means half of News and Common Crawl.

essary. We adopted the data denoise strategy by Wang et al. (2019, 2018). The strategy uses a small amount of high-quality data to tune the base model, and then leverages the differences between the tuned model and the baseline to score bilingual data. The score is calculated based on formula 1.

$$score = \frac{\log P(y|x;\theta_{clean}) - \log P(y|x;\theta_{noise})}{|y|} \quad (1)$$

Where $\theta_{noise}$ denotes the model trained with noisy data; $\theta_{clean}$ denotes the model after fine-tuning on a small amount of clean bilingual data, and $|y|$ denotes the length of the sentence. Higher $score$ means higher quality.

## 3 System Overview

Our method basically follows our previous training strategies (Wei et al., 2020, 2021), such as commonly used Back-Translation (Edunov et al., 2018), Iterative Joint Training (Zhang et al., 2018), Multilingual enhancement (Johnson et al., 2017; Kudugunta et al., 2019; Zhang et al., 2020), Data Diversification (Nguyen et al., 2020) (for details, please refer to our previous work Yang et al. (2021)), Ensemble and Fine-tuning, etc. We will not detail these strategies in this report. The following paper focuses on new strategies used in this year.

### 3.1 Model

We continue using Transformer (Vaswani et al., 2017) as our NMT architecture, but we do not use the four model variants as last year. For convenience, we only use a 25-6 deep model architecture. The parameters of the model are the same as Transformer-big. We just change the post-layer-normalization to the pre-layer-normalization, and increase the encoder layers to 25.

### 3.2 R-Drop

Dropout-like method (Srivastava et al., 2014; Gao et al., 2022) is a powerful and widely used technique for regularizing deep neural networks. Though it can help improve training effectiveness, the randomness introduced by dropouts may lead to inconsistencies between training and inference. R-Drop (Wu et al., 2021) forces the output distributions of different sub models generated by dropout be consistent with each other. Therefore, we use R-Drop training strategy to augment the baseline model for each track and reduce inconsistencies between training and inference.

### 3.3 Pretrained NMT Model

There are many pre-trained Sequence-to-Sequence models, such as Mbart (Liu et al., 2020), MT5 (Xue et al., 2020), M2M-100 (Fan et al., 2020), etc. These pre-trained models are very useful for ultra-low resource tasks. For the ultra-low-resource track Liv↔En, very few bilingual data (1k) is available, so we use a method similar to Adelani and Alabi (2022) to continue training on the basis of M2M-100 (418M) [1]. Since M2M-100 does not support the Liv language, we select an existing language tag (Estonian) similar to Liv to identify this language. For unknown tokens in Liv, we replace them with very low-frequent words in the vocabulary. We find this strategy effective for performance improvement.

### 3.4 Noised Self-Training

Self-training (Imamura and Sumita, 2018) (ST), also known as Forward translation (Wu et al., 2019b), usually refers to using a forward NMT model to translate source-side monolingual data so as to generate synthetic bilinguals, which aims at

[1] https://dl.fbaipublicfiles.com/m2m_100/418M_last_checkpoint.pt

| System | WMT20 | WMT21 | Med20 | Flores | Avg | WMT22 |
|---|---|---|---|---|---|---|
| baseline | 41.6 | 32.2 | 34.3 | 42.2 | 37.6 | - |
| R-Drop | 43.4 | 32.9 | 35.6 | 44.0 | 39.0 | - |
| Data Rejuvenation | 43.5 | 33.0 | 35.4 | 44.3 | 39.5 | - |
| Data Diversification | 44.8 | 33.4 | 35.7 | 44.5 | 39.6 | - |
| ST+BT | 45.0 | 33.8 | 36.6 | 45.0 | 40.1 | 46.0 |
| Finetune & Ensemble (constrain) | - | - | - | - | - | **47.8** |
| Domain Data (unconstrain) | - | - | - | - | - | **49.7** |

Table 2: En→Zh BLEU scores on WMT 2020 News (WMT20), WMT 2021 News (WMT21), WMT 2020 Biomedical (Med20) and Flores test sets, and their average (Avg) scores based on different training strategies. We also report part of WMT 2022 (WMT22) test set results.

| System | WMT20 | WMT21 | Med20 | Flores | Avg | WMT22 |
|---|---|---|---|---|---|---|
| baseline | 28.6 | 23.5 | 26.3 | 30.5 | 27.2 | - |
| R-Drop | 30.4 | 25.0 | 28.3 | 31.8 | 28.9 | - |
| Data Rejuvenation | 31.3 | 26.2 | 28.4 | 31.3 | 29.3 | - |
| Data Diversification | 32.5 | 27.8 | 29.5 | 31.9 | 30.4 | - |
| ST+BT | 33.3 | 28.1 | 29.6 | 32.0 | 30.7 | 26.0 |
| Finetune & Ensemble (constrain) | - | - | - | - | - | **27.7** |
| Domain Data (unconstrain) | - | - | - | - | - | **29.8** |

Table 3: Zh→En BLEU scores on WMT 2020 News (WMT20), WMT 2021 News (WMT21), WMT 2020 Biomedical (Med20) and Flores test sets, and their average (Avg) scores based on different training strategies. We also report part of WMT 2022 (WMT22) test set results.

increasing the training data size. Forward translation usually relies on beam search-based (Freitag and Al-Onaizan, 2017) decoding when generating synthetic data. He et al. (2019) find that drop-out plays an important role in ST and adding a certain noise to the original text can further improve the effect of ST, which is called Noised ST. We adopt this method during training.

### 3.5 Data Rejuvenation

We score all the training bilingual data through Equation 1, and filter out 10% - 20% of the data according to the score distribution. We use the remaining 80% - 90% clean data to continue training on the previous model for denoising. This strategy is particularly effective with noisy data and is used in several several languages in this task. We refer to it as Data Rejuvenation in the following.

## 4 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training and sacreBLEU (Post, 2018) to measure system performances. The main parameters are as follows: Each model is trained using 8 V100 GPUs. The size of each batch is set as 2048, parameter update frequency as 4, and learning rate as 5e-4 (Vaswani et al., 2017). 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 language pairs. R-Drop is used in model training, and we set parameter $\lambda$ to 5 for all language pairs.

## 5 Results and Analysis

### 5.1 Zh↔En

Regarding Zh↔En, we use R-Drop, Knowledge Distillation (Kim and Rush, 2016), Self Training + Back Translation, and fine-tuning. The results of Zh→En and En→Zh are shown in Tables 2 and 3.

To better measure the generalizability of our models, we also calculate BLEU on WMT Biomedical 2020 and Flores test sets (Goyal et al., 2021).

We see that R-Drop can stably bring about 1.5 BLEU improvement, and data enhancement can bring 1.0 BLEU improvement. In the final result we submitted, we only use the news test sets to fine-tune the model, but we see that it was still able to bring 1 BLEU improvement on the WMT 2022 test set.

In the end, our submission uses a combination of our domain-related in-house data and the WMT

| System | WMT20 | WMT21 | Med20 | Flores | Avg | WMT22 |
|---|---|---|---|---|---|---|
| baseline | 22.9 | 26.2 | 32.7 | 30.8 | 28.2 | - |
| ST+BT | 23.8 | 27.9 | 33.1 | 31.3 | 29.0 | - |
| ST+BT+R2L | 24.1 | 28.4 | 32.1 | 31.6 | 29.1 | - |
| Data Rejuvenation | 22.9 | 27.1 | 34.9 | 31.5 | 29.1 | 27.2 |
| Common Crawl | 24.1 | 28.6 | 34.5 | 32.7 | 30.0 | 29.4 |
| Finetune | - | - | - | - | - | 30.4 |
| Ensemble | - | - | - | - | - | **30.8** |

Table 4: En→Ru BLEU scores on WMT 2020 News (WMT20), WMT 2021 News (WMT21), WMT 2020 Biomedical (Med20) and Flores test sets, and their average (Avg) scores based on different training strategies. We also report part of WMT 2022 (WMT22) test set results.

| System | WMT20 | WMT21 | Med20 | Flores | Avg | WMT22 |
|---|---|---|---|---|---|---|
| baseline | 36.1 | 36.7 | 41.1 | 34.1 | 37.0 | - |
| ST+BT | 37.5 | 38.1 | 40.4 | 35.1 | 37.8 | - |
| ST+BT+R2L | 37.7 | 38.4 | 41.4 | 36.2 | 38.4 | 42.8 |
| Data Rejuvenation | 37.1 | 38.1 | 42.7 | 36.7 | 38.7 | 43.0 |
| Common Crawl | 37.4 | 38.1 | 42.6 | 36.5 | 38.7 | 43.4 |
| Finetune | - | - | - | - | - | 44.6 |
| Ensemble | - | - | - | - | - | **45.1** |

Table 5: Ru→En BLEU scores on WMT 2020 News (WMT20), WMT 2021 News (WMT21), WMT 2020 Biomedical (Med20) and Flores test sets, and their average (Avg) scores based on different training strategies. We also report part of WMT 2022 (WMT22) test set results.

data, and we find that domain-related data is critical for quality improvement. By using the extra data, we get an improvement of about 2.0 BLEU over using only the WMT data. Our final Zh→En and En→Zh submissions achieve 49.7 and 29.8 BLEU respectively.

### 5.2 Ru↔En

Regarding Ru↔En (Table 4 and 5), we use strategies including Iterative Self Training + Back Translation, R2L enhancement, and general domain monolingual enhancement.

We see that in addition to the average 1 BLEU improvement brought by fine-tune, the most effective strategy is adding more general domain data. On En→Ru, after the Common Crawl monolingual is added, we observe 2.0 BLEU improvement on WMT 2022 test set.

The data enhancement strategy could bring stable improvement like that in Zh↔En, with an increase of 2 BLEU compared to the baseline model in an average.

The BLEU scores of our final Ru→En and En→Ru submissions are 45.1 and 30.8 respectively.

| System | En→Liv | Liv→En |
|---|---|---|
| M2M-100 finetune | 8.0 | 16.0 |
| OOV process | 9.6 | 17.6 |
| Multilingual | 11.0 | 21.6 |
| Iter Tagged BT | 13.3 | 24.0 |
| Noised ST | 14.6 | - |
| R-Drop | 15.1 | 25.8 |
| WMT22 Submission | **12.8** | **23.4** |

Table 6: The results of Liv↔En for WMT 2022 dev test set. We remove overlapping sentences in the dev set that also appear in the training set.

### 5.3 Liv↔En

Regarding Liv↔En (Table 6), we first fine-tune the M2M-100 model with 1K bilingual data, and then replace the out-of-vocabulary (OOV) token in Liv with low-frequency sub-words in the vocabulary, we see that this strategy brings 1.6 BLEU improvement on En→Liv.

Then we use the Liv/Et and Liv/Lv data together to fine-tune the model. This strategy can bring significant improvement on both directions (1.4 BLEU on En→Liv and 4 BLEU on Liv→En. It should be pointed out that regarding En→Liv, we use additional data from Et→Liv and Lv→Liv, while for

| System | dev | Flores | Avg |
|---|---|---|---|
| R-Drop | 31.5 | 33.2 | 32.4 |
| Data Rejuvenation | 32.1 | 33.5 | 32.8 |
| Sampling BT | 33.2 | 32.9 | 33.1 |
| Finetune | 33.1 | 33.0 | 33.0 |
| Ensemble | 33.2 | 33.4 | 33.3 |
| WMT22 Submission | | **18.1** | |

Table 7: The results of En→Hr on WMT 2022 dev test set and Flores.

Liv→En, we use data from Liv→Et and Liv→Lv to enhance the model.

We do three rounds of Tagged BT (Caswell et al., 2019) in total and observe that the improvement is still significant (an average improvement of 3 BLEU on two directions). For En→Liv, we adopt the strategy of Noised ST because we have a large amount of English monolinguals. We used 1M English monolinguals for Noised ST. We see that this strategy can bring an additional 1.3 BLEU improvement.

Additionally, we employ the R-Drop strategy during training and find that on Liv2En, this strategy brings an improvement of 1.8 BLEU.

Finally, using dev fine-tune and ensemble of 4 models, our submissions achieve 12.8 BLEU on En→Liv, and 23.4 BLEU on Liv→En.

### 5.4 En→Hr

The results of En→Hr are shown in Table 7. We use 22M Hr monolinguals for BT and find that the results on the dev set is different from that on the test set as the magnitude of improvements are inconsistent. The overall improvement on dev set is only 0.8 BLEU, but 3 BLEU on the test set. The main improvement is brought by data denoising. We assume that this is because the provided En2Hr bilingual data is highly noisy. Our final submission achieves 18.1 BLEU.

### 5.5 Uk↔En and Cs↔Uk

Regarding Uk↔En (Table 8), we conduct Sampling BT and see 2.2 BLEU improvement on Uk→En but no improvement on En↔Uk. After adding self-training data, an additional 0.5 BLEU improvement is gained on Uk→En. We then use real bilinguals data to continue training the model that have been augmented with synthetic data. This strategy further leads to an average improvement of 0.4 BLEU. We do not use dev fine-tuning but directly ensemble the 4 models. The final En→Uk and Uk→En submissions achieve 26.5 and 41.6 BLEU respectively on the WMT22 test set.

The strategy for Cs↔Uk is basically the same as that for Uk↔En, but we further apply multilingual enhancement. We use additional En→Uk data for enhancing Cs→Uk translation and En→Cs data for enhancing Uk→Cs translation. Multilingual enhancement brings 1.2 BLEU improvement on Uk→ Cs. Monolingual data augmentation also brings significant improvement. Ensemble further leads to 1 BLEU increase on Uk→Cs. Our final Cs↔Uk submissions achieve 36.0 BLEU on the WMT22 test sets.

## 6 Discussion

### 6.1 General Domain

In this year, WMT changed its focus on news domain to the broader general task, with three additional domains putting into consideration (social, conversational, and ecommerce). We also use test sets from other domains to measure the generalizability of our models.

However, for language pairs we participate in, most of the knowledge in domains other than news can only be learned from Common Crawl monolinguals. Without in-domain data, a model's performance in social, conversational and ecommerce domains can hardly be improved. We add additional bilingual data related to the three domains for the Zh↔En track and observe an average of 2.0 BLEU improvement. As a result, how to maximize the effectiveness of in-domain data is crucial.

### 6.2 Evaluation Method

N-gram matching metrics such as BLEU and chrF (Popović, 2015) are widely used in machine translation evaluation. However, as machine translation technology improves, relying only on BLEU to evaluate a model's performance become increasingly risky. For example, in last year's evaluation, the BLEU score of our De→En model ranks among the top, but the human evaluation results show that our model performs the worst. In this year's En→Uk evaluation, widely-used back-translation lead to no BLEU increase as shown in Table 8. So far, we are not sure whether back-translation does lead to no improvement or the improvement cannot be measured by BLEU. We believe that more researches are required on robust metrics (Sellam et al., 2020; Rei et al., 2020), reliable test set constructions, and sound human evaluation methods

| System | En→Uk | Uk→En | Cs→UK | Uk→Cs |
|---|---|---|---|---|
| baseline | 31.7 | 38.7 | 24.1 | 22.3 |
| Multilingual | - | - | 24.6 | 23.5 |
| Sampling BT | 31.7 | 40.9 | 25.7 | 24.2 |
| ST + BT | 31.5 | 41.4 | 25.4 | 23.9 |
| Data Rejuvenation | 31.9 | 41.8 | 25.7 | 24.2 |
| Ensemble | 32.9 | 41.9 | 26.3 | 25.1 |
| WMT22 Submission | **26.5** | **41.6** | **36.0** | **36.0** |

Table 8: The results of Uk↔En and Uk↔Cs for WMT 2022 dev set.

considering the great advances in NMT and subtle differences among systems.

## 7 Conclusion

This paper presents the submissions of HW-TSC to the WMT 2022 General Machine Translation Task. We participate in six language pairs and perform experiments with a series of pre-processing and training strategies. The effectiveness of each strategy is demonstrated. Our experiments show that in very low-resource scenarios, fine-tuning on pre-trained NMT models can significantly improve system performance. R-Drop also brings stable improvement across languages. Certainly, commonly-used data augmentation strategies are still effective for model training. Our submissions finally achieve competitive results in the evaluation.

## References

David Adelani and Jesujoba et al. Alabi. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. 2017. Ensemble distillation for neural machine translation. *CoRR*, abs/1702.01802.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation.

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation.

Kenji Imamura and Eiichiro Sumita. 2018. Nict self-training approach to neural machine translation at nmt-2018. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 110–115.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019. The niutrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 257–266.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems*, volume 33, pages 10018–10029. Curran Associates, Inc.

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. *arXiv preprint arXiv:1904.01038*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 374–381.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019. Dynamically composing domain-data selection with clean-data selection by "co-curricular learning" for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1282–1292, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. pages 133–143.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 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, Online. Association for Computational Linguistics.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. HW-TSC's participation in the WMT 2020 news translation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 293–299, Online. Association for Computational Linguistics.

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.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019a. Exploiting monolingual data at scale for neural machine translation.

In *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, Hong Kong, China, November 3-7, 2019*, pages 4205–4215.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019b. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, Chuanfei Xu, Min Zhang, and Ying Qin. 2021. HW-TSC's submissions to the WMT21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884, Online. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Thirty-Second AAAI Conference on Artificial Intelligence*.