HW-TSC at IWSLT2023: Break the Quality Ceiling of Offline Track via Pre-Training and Domain Adaptation

Zongyao Li, Zhanglin Wu, Zhiqiang Rao, Xie YuHao, Guo JiaXin,
Daimeng Wei, Hengchao Shang, Wang Minghan, Xiaoyu Chen
Zhengzhe YU, Li ShaoJun, Lei LiZhi, Hao Yang
Huawei Translation Service Center, Beijing, China
{lizongyao,wuzhanglin2,raozhiqiang,xieyuhao2,guojiaxin1,
weidaimeng,shanghengchao,wangminghan,chenxiaoyu35,
yuzhengzhe,lishaojun18,leilizhi,yanghao30}@huawei.com

Abstract

This paper describes HW-TSC's submissions to the IWSLT 2023 Offline Speech Translation task, including speech translation of talks from English to German, English to Chinese and English to Japanese. We participated in all three tracks (Constrained training, Constrained with Large Language Models training, Unconstrained training), with using cascaded architectures models. We use data enhancement, pre-training models and other means to improve the quality of ASR, and use a variety of techniques including R-Drop, deep model, domain data selection, etc. to improve the quality of NMT. Compared with last year's best results, we have improved by 2.1 BLEU in the MuST-C English-German test set.

1 Introduction

The goal of the Offline Speech Translation Task is to examine automatic methods for translating audio speech in one language into text in the target language. In recent years, end-to-end system and cascade system are fundamental pipelines for speech translation tasks. Traditional cascade system is comprised of continuing parts, automatic speech recognition (ASR) is responsible for generating transcripts from audios and machine translation (MT) model aims at translating ASR outputs from source language into target language. ASR model like Conformer (Gulati et al., 2020) and S2T-Transformer (Synnaeve et al., 2019) are commonly used. MT models like Transformer (Vaswani et al., 2017) can be considered as a standard configuration. The End-to-end systems use a model to directly recognize speech into target text in another language.

The cascade system will cause some "missing information" due to the two encoding and decoding processes of ASR and MT. At the same time, the disadvantage of the end-to-end system is the lack of sufficient training data. However, with a fully trained cascade system, the accuracy of ASR and MT will reach a higher level. So from the results, the BLEU of the cascaded system will be higher than that of the end-to-end system. Currently in the industry, the mainstream speech translation system is still based on the cascade system. We use the cascade system for this task, mainly to further improve the performance of speech translation.

In this work, we carefully filter and preprocess the data, and adopt various enhancement techniques, such as pre-training model, data enhancement, domain adaptation, etc., to optimize the performance of ASR. We build machine translation systems with techniques like back translation (Edunov et al., 2018), domain adaptation and R-drop (Wu et al., 2021), which have been proved to be effective practices.

The main contribution of this paper can be summarized as follows:

1) According to the characteristics of three different tracks (constrained, constrained with large language models (LLM), and unconstrained), we use different strategies to optimize the results of ASR. After careful fine-tuning, the WER of the ASR system of the three tracks have achieved good performance.

2) Explored the multilingual machine translation model, and tried a variety of model enhancement strategies, and finally achieved good results on the MUST-C test set.

Section 2 focuses on our data processing strategies while section 3 describes the training techniques of ASR, including model architecture and training strategy, etc. Section 4 describes the training techniques of MT, and section 5 presents our experiment results.

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 187–193
July 13-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Duration(h)
LibriSpeech	960
MuST-C	590
CoVoST	1802
TEDLIUM3	453
Europarl	161
VoxPopuli	1270

Table 1: Data statistics of our ASR corpora.

2 Datasets and Preprocessing

2.1 ASR Data

There are six different datasets used in the training of our ASR models, such as MuST-C V2 (Cattoni et al., 2021), LibriSpeech (Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), CoVoST 2(Wang et al., 2020), VoxPopuli (Wang et al., 2021), Europarl-ST (Iranzo-Sánchez et al., 2020), as described in Table 1. We use the exactly same data processing strategy to train our ASR models following the configuration of (Wang et al., 2022). We extend one data augmentation method (Zhang et al., 2022): adjacent voices are concatenated to generate longer training speeches. Tsiamas et al. (2022) propose Supervised Hybrid Audio Segmentation (SHAS), a method that can effectively learn the optimal segmentation from any manually segmented speech corpus. For test set, we use SHAS to split long audios into shorter segments.

2.2 MT Data

We used all provided data, including text-parallel and speech-to-text-parallel, text-monolingual data, and use the exactly same data processing strategy to process our MT data following (Wei et al., 2021). Data sizes before and after cleaning are listed in Table 2.

3 ASR Model

3.1 Constrained training

In this track, we trained the constrained ASR model using the Conformer (Gulati et al., 2020) and U2 (Zhang et al., 2020b) model architectures. The first model is standard auto-regressive ASR models built upon the Transformer architecture. The last one is a unified model that can perform both streaming and non-streaming ASR, supported by the dynamic chunking training strategy. The model configurations are as follows:

1) **Conformer**: The encoder is composed of 2 layers of VGG and 16 layers of Conformer, and the decoder is composed of 6 layers of Transformer. The embedding size is 1024, and the hidden size of FFN is 4096, and the attention head is 16.

2) **U2**: Two convolution subsampling layers with kernel size 3*3 and stride 2 are used in the front of the encoder. We use 12 Conformer layers for the encoder and 6 Transformer layers for the decoder. The embedding size is 1024, and the hidden size of FFN is 4096, and the attention head is 16.

During the training of ASR models, we set the batch size to the maximum of 20,000 frames percard. Inverse sqrt is used for lr scheduling with warm-up steps set to 10,000 and peak lr set as 5e-4. Adam is used as the optimizer. All ASR models are trained on 8 A100 GPUs for 100 epochs. Parameters for last 5 epochs are averaged. Audio features are normalized with utterance-level CMVN for Conformer, and with global CMVN for U2. All audio inputs are augmented with spectral augmentation (Park et al., 2019), and Connectionist Temporal Classification (CTC) is added to make models converge better.

3.2 Constrained with Large Language Models training

Large Language Models (LLM) is currently the mainstream method in the field of artificial intelligence. In ASR, the pre-training model has been proved to be an effective means to improve the quality, especially the models such as wav2vec (Schneider et al., 2019) and Hubert (Hsu et al., 2021) have been proposed in recent years. Li et al. (2020) combine the encoder of wav2vec2 (Baevski et al., 2020) and the decoder of mBART50 (Tang et al., 2020) to fine-tune an end2end model. We also adopt a similar strategy, but combine the encoder of wav2vec2 and the decoder of mBART50 to fine-tune an ASR model (w2v2-mBART). Due to the modality mismatch between pre-training and fine-tuning, in order to better train cross-attention, we freeze the self-attention of the encoder and decoder. We first use all the constrained data for fine-tuning, and only use the MUST-C data after 30 epochs of training.

3.3 Unconstrained training

Whisper (Radford et al., 2022) is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. It show that

language pairs	Raw Data	Filter Data	LaBSE Filter Data	Domain Selection
En2De	19.8M	14.5M	5.8M	0.4M
En2Zh	8.1M	5.5M	2.2M	0.4M
En2Ja	16.4M	14.1M	5.6M	0.4M

Table 2: Bilingual data sizes before and after filtering used in tasks.

the use of such a large and diverse dataset leads to improved robustness to accents, background noise and technical language. The Whisper architecture is a simple end-to-end approach, implemented as an encoder-decoder Transformer. Even though it enables transcription in multiple languages, we only use its speech recognition feature, transcribing audio files to English text. In this task, we use it as a pre-trained model, and use the MUST-C dataset for fine-tuning to improve its performance in specific domains. We trained for 2 epochs with a small learning rate of 10e-6.

4 Neural Machine Translation

4.1 Model architecture

Transformer is the state-of-the-art model in recent machine translation evaluations. There are two parts of research to improve this kind: the first part uses wide networks (eg: Transformer-Big), and the other part uses deeper language representations (eg: Deep Transformer (Wang et al., 2017, 2019a)). Under the constrained conditions, we combine these two improvements, adopt the Deep Transformer-Big model structure, and train a one-to-many multilingual NMT model (Johnson et al., 2017; Zhang et al., 2020a) from scratch using bilingual data of three language pairs (En2De, En2Zh, En2Ja) provided by the organizers. The main structure of Deep Transformer-Big is that it features pre-layer-normalization and 25-layer encoder, 6-layer decoder, 16-head self-attention, 1024-dimensional embedding and 4096-dimensional FFN embedding.

We trained the constrained model using all the provided data, and trained the unconstrained model with the WMT data. But after domain adaptation, the performance of the two is similar. Therefore, in this task, we only use the constrained MT model.

4.2 Multi-stage Pre-training

In order to get a better model effect, we optimize the model in several stages. First, we use the data of all three language pairs to train a one-to-many multilingual model, and add tags (<ja>, <zh>, <de>) at the beginning of the source sentence respectively. Second, use LaBSE (Feng et al., 2020) to filter the bilingual data, and use the filtered data for incremental training. In Table 2, there are the number of filtered data for each languages. Then, for the three languages, the backward models are trained separately, and the monolingual datas are used for backward translation (BT). Finally, we combine backward translation and forward translation (FT) for iterative joint training (Zhang et al., 2018). After the above several stages, a base model with better performance is obtained, which can be used for further optimization.

4.3 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 base model for each track and reduce inconsistencies between training and inference.

4.4 Domain Adaptation

Since the quality of the translation model is easily affected by the domain, we try to select domain-related data to incrementally train the model. We adopted the domain adaptation strategy by (Wang et al., 2019b). The strategy uses a small amount of in-domain data to tune the base model, and then leverages the differences between the tuned model and the base to score bilingual data. The score is calculated based on formula 1.

$$score = \frac{logP(y|x;\theta_{in}) - logP(y|x;\theta_{base})}{|y|} \quad (1)$$

Where θ_{base} denotes the base model; θ_{in} denotes the model after fine-tuning on a small amount of in-domain data, and |y| denotes the length of the sentence. Higher score means higher quality.

System	En2De	En2Ja	En2Zh
Constrained	37.28	20.26	28.91
Constrained with LLM	37.96	20.29	28.91
Unconstrained	38.71	20.34	28.93

Table 3: The BLEU of speech translation on tst-COM.

System	tst-COM	tst2018	tst2019	tst2020	avg
Conformer	5.3	9.3	6.7	8.9	7.6
U2	6.1	9.8	6.6	8.7	7.8
w2v2-mBART	4.9	9.3	6.9	8.9	7.5
Whisper	4.5	11.0	5.4	6.6	6.8
Whisper fine-tuning	4.3	8.5	6.3	7.9	6.8

Table 4: The experimental results of ASR. We present WER performance of tst-COM, tst2018, tst2019 and tst2020.

System	En2De	En2Ja	En2Zh
One2Many	36.22	15.43	29.05
+ LaBSE bitext	37.58	15.48	29.48
+ Domain adaptation	41.55	17.08	29.27
+ Iter FTBT	43.03	17.86	29.82
+ Dev fine-tuning	43.66	20.88	30.48

Table 5: The BLEU of MT using tst-COM golden transcription.

System	En2De	En2Ja	En2Zh
One2Many	31.54	14.08	26.69
+ LaBSE bitext	32.65	13.88	27.14
+ Domain adaptation	35.96	15.4	27.15
+ Iter FTBT	36.38	15.81	27.98
+ Dev fine-tuning	37.83	18.6	28.86
+ Robustness	38.71	20.34	28.93

Table 6: The BLEU of MT using tst-COM transcription by the Whisper fine-tuning model.

In this task, we use TED and MUST-C data as in-domain data. We score all the training bilingual data through Equation 1, and filter out 80% - 90% of the data according to the score distribution. We use the remaining 0.4M in-domain data to continue training on the previous model.

4.5 Robustness to ASR Noise

We use two methods to improve the robustness of the system to ASR output noise.

Synthetic Noise Generation. We refer to the method proposed in Guo et al. (2022) to synthesize part of the noise data to enhance the robustness of the model.

ASR Transcript Data. Because some triplet data are provided in this task, including $audio$, $source$ and $target$. We use the trained ASR to transcribe the audio file to get $source'$, and finally get the MT training data like $(source', target)$. The $source'$ transcribed by ASR may have some errors, but when used in MT, it will increase the robustness of the MT encoder.

When using the data generated above, we refer to the tagged BT method (Caswell et al., 2019), and add a special token at the beginning of the source sentence.

5 Experiments and Results

We use the open-source fairseq (Ott et al., 2019) for training, word error rate (WER) to evaluate the ASR models and report case-sensitive SacreBLEU (Post, 2018) scores for machine translation. We evaluated our system on the test sets of MuST-C tst-COMMON (tst-COM).

Table 3 is our results on three languages for three tracks (Constrained, Constrained with LLM, Unconstrained). After a series of optimizations, although the ASR results of the three systems are somewhat different, the BLEU of all systems are very close. Since there is no testset for iwslt2022, we only compared with last year's teams on tst-COM. Compared with last year's best results (Zhang et al., 2022), we have improved by 2.1 BLEU in the MuST-C En2De test set; in En2Zh and En2Ja, we have achieved close to last year's best results.

We analyze the main reasons for the similar results of the three systems: 1. The three systems use the same MT, and our MT system has the ability to correct wrong input after the robustness is enhanced. 2. Using the same data to finetuning the three ASR systems, the WER are relatively close.

5.1 Automatic Speech Recognition

We compare the results of different model architectures, the overall experimental results about ASR is described in Table 4. We evaluated our system on the test sets of tst-COM, IWSLT tst2018/tst2019/tst2020 respectively. For long audio in the test set, we use SHAS for segmentation. We calculate the WER after the reference and hypothesis are lowercased and the punctuation is removed.

In Table 4, all ASR systems achieve good performance, and the results are relatively close. Conformer and U2 are trained using constrained data. w2v2-mBART is obtained through fine-tuning using pre-trained models, which are constrained. Whisper is the result of transcribing long audio without segmentation using the native whisper medium model. Whisper fine-tuning is obtained after fine-tuning on MuST-C dataset, with using the Whisper medium model. The WER of Conformer and U2 is relatively close. In submitting the results of constrained track, we use Conformer as the final ASR system. The experimental results show that pre-trained models exhibit their advantages, w2v2-mBART can achieve better results than just training with constrained data. Whisper itself has a very good performance in the general domain, and after fine-tuning, it has even better results in the specific domain. However, it is very difficult to perform finetuning on whisper and improve the performance of all domains. WER performance on tst2019 and tst2020 has deteriorated.

5.2 Neural Machine Translation

We evaluate the performance of the MT model in detail on the MUST-C test set. Table 5 shows the performance results of each optimization strategy using golden as the source; Table 6 uses the transcription generated by Whisper fine-tuning model as the source. The results show that there is a gap in BLEU between golden and transcription of ASR, which is mainly due to errors (punctuation, capitalization, vocabulary, etc.) in transcription of ASR. On the En2De test set, this gap is particularly wide.

One2Many is a multilingual model trained using the R-drop strategy, and has achieved relatively good performance on the test set. LaBSE can bring a little improvement to the model, and domain adaptation can bring a huge improvement to the model, which proves the effectiveness of our strategy. Iterative joint training with FT and BT (Iter FTBT) is also an effective mean to improve quality. After dev fine-tuning, the results are already very competitive. With improving the robustness of the system to ASR output, our BLEU in En2De, En2Zh, and En2Ja are 38.71, 20.34, and 28.93, respectively.

6 Conclusion

This paper presents our offline speech translation systems in the IWSLT 2023 evaluation. We explored different strategies in the pipeline of building the cascade system. In the data preprocessing, we adopt efficient cleansing approaches to build the training set collected from different data sources. We tried various ASR training strategies and achieved good performance. For the MT system, we have used various methods such as multilingual machine translation, R-drop, domain adaptation, and enhanced robustness. Finally, compared with last year's best results, we have improved by 2.1 BLEU in the MuST-C English-German test set.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2022. Bi-simcut: A simple strategy for boosting neural machine translation. *arXiv preprint arXiv:2206.02368*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al.

2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Bao Guo, Mengge Liu, Wen Zhang, Hexuan Chen, Chang Mu, Xiang Li, Jianwei Cui, Bin Wang, and Yuhang Guo. 2022. The xiaomi text-to-text simultaneous speech translation system for iwslt 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 216–224.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Roselló, Adria Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

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.

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

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*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

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.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ioannis Tsiamas, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv preprint arXiv:2202.04774*.

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.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, et al. 2022. The hw-tsc's simultaneous speech translation system for iwslt 2022 evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 247–254.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. 2017. Deep neural machine translation with linear associative unit. *arXiv preprint arXiv:1705.00861*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019a. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*.

Wei Wang, Isaac Caswell, and Ciprian Chelba. 2019b. Dynamically composing domain-data selection with clean-data selection by" co-curricular learning" for neural machine translation. *arXiv preprint arXiv:1906.01130*.

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.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. 2020b. Unified streaming and non-streaming two-pass end-to-end model for speech recognition. *arXiv preprint arXiv:2012.05481*.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, et al. 2022. The ustc-nelslip offline speech translation systems for iwslt 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 198–207.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.