Improving the Quality of IWLST 2024 Cascade Offline Speech Translation and Speech-to-Speech Translation via Translation Hypothesis Ensembling with NMT models and Large Language Models

Zhanglin Wu, JiaXin Guo, Daimeng Wei, Zhiqiang Rao,
Zongyao Li, Hengchao Shang, Yuanchang Luo, Li ShaoJun, Hao Yang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,guojiaxin1,weidaimeng,raozhiqiang,
lizongyao,shanghengchao,luoyuanchang1,lishaojun18,yanghao30}@huawei.com

Abstract

This paper presents HW-TSC's submission to the IWSLT 2024 Offline Speech Translation Task and Speech-to-Speech Translation Task. The former includes three translation directions: English to German, English to Chinese, and English to Japanese, while the latter only includes the translation direction of English to Chinese. We attend all three tracks (Constraint training, Constrained with Large Language Models training, and Unconstrained training) of offline speech translation task, using the cascade model architecture. Under the constrained training track, we train an ASR model from scratch, and then employ R-Drop and domain data selection to train the NMT model. In the constrained with Large Language Models training track, we use Wav2vec 2.0 and mBART50 for ASR model training initialization, and then train the LLama2-7B-based MT model using continuous training with sentence-aligned parallel data, supervised fine-tuning, and contrastive preference optimization. In the unconstrained training track, we fine-tune the whisper model for speech recognition, and then ensemble the translation results of NMT models and LLMs to produce superior translation output. For the speech-to-speech translation Task, we initially employ the offline speech translation system described above to generate the translated text. Then, we utilize the VITS model to generate the corresponding speech and employ the OpenVoice model for timbre cloning.

1 Introduction

Recent advances in deep learning allow us to address traditional NLP tasks in a new and significantly different manner. One such task is speech translation, involving automatic speech recognition (ASR) (Gulati et al., 2020) system and machine translation (MT) (Vaswani et al., 2017) system. Another task is speech-to-speech translation (S2S), which involves ASR system, MT system, and text-to-speech (TTS) (Ren et al., 2020) system. Recent trends tend to utilize a single neural network to directly translate input speech from one language to text or speech in another language, bypassing intermediate symbolic representations. The results shows that the performance of end-to-end models is nearing that of cascade solutions, but the effectiveness comparison between the two technologies remains unclear. Both methods face specific challenges. The primary challenge with the end-to-end approach is the lack of training data, while the cascade method has to go through the ASR, MT and even TTS processes, leading to the errors accumulation. Due to the data insufficiency in end-to-end training, We ultimately chose the cascade approach on the IWSLT 2024 offline speech translation task and speech-to-speech translation task.

For the IWSLT offline speech translation task, we apply different training strategies across the three tracks, adapting to diverse data and model conditions. In the constrained training track, we initiate training with an ASR model from scratch, followed by the utilization of R-Drop (Wu et al., 2021) and domain data selection (Wang et al., 2019b) techniques to train the NMT model. Within the constrained with Large Language Models (LLMs) training track, we commence ASR model training initialization using Wav2vec 2.0 (Baevski et al., 2020) and mBART50 (Tang et al., 2020). Subsequently, we train the LLama2-7B-based (Touvron et al., 2023) MT model through continual pre-training with sentence-aligned parallel data (Guo et al., 2024), supervised fine-tuning (Xu et al., 2023), and contrastive preference optimization (CPO) (Xu et al., 2024). In the unconstrained training track, we fine-tune the whisper model (Radford et al., 2023) for speech recognition, and then ensemble (Farinhas et al., 2023) the translation outputs of NMT models and LLMs to generate superior translation result. For the IWSLT S2S translation task, we initially employ the offline speech translation system described above to

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 94–100

generate the translated text. Next, we utilize the VITS (Kim et al., 2021) model to generate the corresponding speech and employ the OpenVoice (Qin et al., 2023) model for timbre cloning.

In comparison to last year, our cascade offline speech translation system and S2S translation system is performing significantly better, particularly following translation hypothesis ensembling with NMT models and LLMs.

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. In the test phase, we use SHAS to split long audios into shorter segments.

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

Table 1: Data statistics of ASR corpus.

2.2 MT Data

We use the same data processing strategy following (Wu et al., 2023) to extract our MT data from the officially available text-parallel and speech-to-text-parallel data. Table 2 illustrates the bilingual data sizes after labse filtering (Feng et al., 2022) and domain selection (Wang et al., 2019b).

language pairs	en2de	en2ja	en2zh
Clean Data	5.8M	5.6M	2.2M
Domain Data	0.4M	0.4M	0.4M

Table 2: Bilingual data sizes of MT corpus.

3 ASR Model

3.1 Constrained training

In this track, we train the constrained ASR model using the Conformer (Gulati et al., 2020) and U2 (Zhang et al., 2020) 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 NPUs 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 the model converge better.

3.2 Constrained with LLMs training

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., 2023) 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 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 MT Model

4.1 Constrained training

Transformer stands as the state-of-the-art model in recent machine translation evaluations. Research to enhance this model type is divided into two main avenues: one focuses on using wider networks (e.g., Transformer-Big) (Vaswani et al., 2017), while the other emphasizes deeper language representations (e.g., Deep Transformer (Wang et al., 2017, 2019a)). Under the constrained conditions, we combine these two improvements, adopt the Deep Transformer-Big model structure, and utilize the clean bilingual data filtered by the labse model (Feng et al., 2022) to train the NMT model from scratch. The primary features of Deep Transformer-Big include pre-layer normalization, a 25-layer encoder, a 6-layer decoder, 16-head self-attention, 1024-dimensional embedding, and 4096-dimensional FFN embedding.

To regularize the training of NMT and alleviate the inconsistency between training and inference caused by the randomness of dropout(Srivastava et al., 2014; Gao et al., 2022), we introduce R-Drop(Wu et al., 2021), which forces the output distributions of different sub-models generated by dropout to be consistent with each other.

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.

Specifically, 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.

In the training of NMT models, each model undergoes training utilizing 8 NPUs. The batch size remains fixed at 6144, the update frequency is 2, the dropout is 0.1, and the learning rate is maintained at 5e-4. A total of 4000 warmup steps are executed, and the model is saved every 2000 steps. Additionally, λ is set to 5 for R-Drop.

4.2 Constrained with LLMs training

Generative LLMs have made significant strides in various NLP tasks. However, these advancements have not fully translated to translation tasks, particularly for medium-sized models, which still trail behind traditional supervised encoder-decoder translation models. Previous studies have attempted to enhance the translation ability of these LLMs through prompt translation (Zhang et al., 2023; Moslem et al., 2023), but the improvements remain limited. Fortunately, recent research is making more progress through supervised fine-tuning (SFT) (Zeng et al., 2024), and showing that it is possible to break away from the reliance on massive amounts of parallel data that traditional translation models typically require.

Translate this from [source language] to [target language]:
[source language]: <source_sentence>
[target language]:

Figure 1: The translation prompt used for training and evaluation. [source language] and [target language] represent the full name of the language written in English format, e.g., Translate this from English to Chinese.

Among the officially designated LLMs, we opt to perform MT tasks based on the Llama2-7B base model. To enhance the cross-lingual capability of Llama2-7B, we first adopt the method of continual pre-training with sentence-aligned parallel data (Guo et al., 2024). We construct the data for this format from the clean data listed in Table 2.

Since Guo et al. discovered that constructing translation instruction written in the source language notably improves performance. We then use the domain data to construct a dataset of translation instructions in English format, and leverage this source-language consistent instruction for SFT. The translation prompt used for training and evaluation is shown in Figure 1.

Finally, we introduce CPO (Xu et al., 2024), which trains the model to avoid producing adequate but imperfect translations. To generate the triplet data, we additionally fine-tune a relatively small LM (BLOOM (Shoeybi et al., 2019)) and generate the output for each instance using a simple sampling strategy. With examples of correct and incorrect translations, the model is optimized to distinguish high-quality translations.

During the fine-tuning of LLMs, We adopt LoRA (Hu et al., 2021) method to fine-tune the LLM on 8 NPUs. The epoch size is 1, the batch size is 128, the maximum text length is 512, and the learning rate is 2e-3. Additionally, the weight decay is 0.01.

4.3 Unconstrained training

LLMs are becoming a one-fits-many solution, but they sometimes hallucinate or produce unreliable output. In the unconstrained track, we utilize translation hypothesis ensembling with NMT models and LLMs (Farinhas et al., 2023). First, we gather translation hypotheses from various NMTs and LLMs. Next, we utilize the external model COMET (Rei et al., 2022) to select the optimal result. This involves calculating the average COMET score between each translation hypothesis and the other hypotheses to determine its quality score. Subsequently, we choose the translation hypothesis with the highest quality score as the best result.

5 TTS Model

Several recent end-to-end TTS models enabling single-stage training and parallel sampling have been proposed, but their sample quality does not match that of two-stage TTS systems. VITS (Kim et al., 2021) is a parallel end-to-end TTS method that generates more natural sounding audio than current two-stage models. The method adopts variational inference augmented with normalizing flows and an adversarial training process, which improves the expressive power of generative modeling. In the S2S translation system, we first use the speech translation system to generate the translation text, and then use the VITS model to generate the corresponding speech.

To improve the similarity of synthesized audio's timbre to that of the source language audio, we also use OpenVoice (Qin et al., 2023) model for timbre cloning. It is a versatile voice cloning approach that requires only a short audio clip from the reference speaker to replicate their voice and generate speech in multiple languages.

6 Experiments and Results

The only difference between our S2S translation system and speech translation system is the addition of TTS and timbre cloning modules. Since we did not perform additional training on these two modules, we only present the experimental results of the speech translation system.

We utilize the open-source fairseq (Ott et al., 2019) for training the NMT model, the open-source ALMA (Xu et al., 2023) for fine-tuning LLM model. We assess the ASR models using the word error rate (WER) and evaluate the MT models using case-sensitive SacreBLEU (Post, 2018) and COMET scores. Our ASR system is evaluated on the test sets of tst-COM, while our MT system is evaluated on the test sets of tst-COM and tst2022.

Table 3 presents our final evaluation results for three language pairs across the constrained training, constrained with LLM training, and unconstrained training tracks. As the final evaluation result shows,

	en2de		en2ja		en2zh	
Cascade System	BLEU	COMET	BLEU	COMET	BLEU	COMET
Constrained	**33.64**	0.7762	**19.19**	0.7992	**34.77**	0.8046
Constrained with LLMs	22.55	0.7646	15.70	0.8253	32.66	0.8230
Unconstrained	33.18	**0.7925**	18.46	**0.8325**	33.76	**0.8358**

Table 3: BLEU and COMET of speech translation on tst-2022 test set.

the cascade system based on the NMT model perform better in the BLEU metric, while the cascade system based on the LLM model perform better in the COMET metric. When ensembling the translation results of both NMT and LLM, the cascade system is performing well in both BLEU and COMET.

6.1 ASR Results

We compare the results of different model architectures, the overall experimental results about ASR is described in Table 4. We evaluated our system on tst-COM test set. 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.

ASR System	**tst-COM**
Conformer	5.3
U2	6.1
w2v2-mBART	4.9
Whisper	4.5
Whisper fine-tuning	**4.3**

Table 4: WER of ASR on tst-COM test set.

6.2 MT Results

When evaluating the MT model, we use the Whisper fine-tuning model transcription results as the source text. Since the NMT model performs well on BLEU, we are using BLEU to evaluate the performance of the NMT model at each stage on the tst-COM test set. While the LLM model performs well on COMET, we are using COMET to evaluate the performance of the LLM model at each stage on the tst-2022 test set.

Table 5 is illustrating the BLEU of the NMT model being trained in each phase on the tst-COM test set. These results highlight the importance of employing the domain data selection method to carefully choose domain-specific data for further fine-tuning the model to facilitate domain adaptation. Following this, we utilize tst-dev as a more precise domain dataset for additional fine-tuning, resulting in even greater quality improvements.

NMT System	**en2de**	**en2ja**	**en2zh**
R-Drop baseline	32.65	13.88	27.14
+ Domain data selection	36.33	16.42	27.48
+ tst-dev fine-tuning	38.12	20.05	28.86

Table 5: BLEU of NMT model on tst-COM test set.

Table 6 shows the COMET of the LLM model fine-tuning at each stage on the tst-2022 test set. From the results, it becomes evident that the three methods of continuous training with Interlinear Text Format Documents, SFT, and CPO are orthogonal and can all improve the machine translation capabilities of LLM.

LLM System	**en2de**	**en2ja**	**en2zh**
Llama2-7B	0.5966	0.6925	0.6934
+ continual pre-training	0.7555	0.8016	0.8141
+ SFT	0.7641	0.8150	0.8220
+ CPO	0.7646	0.8253	0.8230

Table 6: COMET of LLM model on tst-2022 test set.

7 Conclusion

This paper presents our cascade speech translation system and S2S translation system in the IWSLT 2024 evaluation. We try several ASR model training strategies and achieve good performance. For the MT system, we explore two research directions based on NMT and LLM, and enhanced them through various technical means. Finally, we achieve further improvements by ensembling the translation results of NMT models and LLMs. For the TTS, we directly use open source models to generate speech and timbre clones. Our experimental results show that LLM-based ASR and MT are promising research directions.

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.

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.

António Farinhas, José GC de Souza, and André FT Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. *arXiv preprint arXiv:2310.11430*.

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.

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*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

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*.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *24th Annual Conference of the European Association for Machine Translation*, page 227.

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*.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

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

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

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.

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*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

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, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, et al. 2023. Treating general mt shared task as a multi-domain adaptation problem: Hw-tsc's submission to the wmt23 general mt shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 170–174.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, pages 19488–19496.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Binbin Zhang, Di Wu, Zhuoyuan Yao, Xiong Wang, Fan Yu, Chao Yang, Liyong Guo, Yaguang Hu, Lei Xie, and Xin Lei. 2020. 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.