The HW-TSC's Speech to Speech Translation System for IWSLT 2022

Jiaxin Guo[1], Yinglu Li[1], Minghan Wang[1], Xiaosong Qiao[1], Yuxia Wang[2], Hengchao Shang[1], Chang Su[1], Yimeng Chen[1], Min Zhang[1], Shimin Tao[1], Hao Yang[1], Ying Qin[1]
[1]Huawei Translation Services Center, Beijing, China
[2]The University of Melbourne, Melbourne, Australia
{guojiaxin1,liyinglu,wangminghan,qiaoxiaosong,
shanghengchao, suchang8,chenyimeng,zhangmin186,
taoshimin,yanghao30,qinying}@huawei.com
yuxiaw@student.unimelb.edu.au

Abstract

The paper presents the HW-TSC's pipeline and results of Offline Speech to Speech Translation for IWSLT 2022. We design a cascade system consisted of an ASR model, machine translation model and TTS model to convert the speech from one language into another language(en-de). For the ASR part, we find that better performance can be obtained by ensembling multiple heterogeneous ASR models and performing reranking on beam candidates. And we find that the combination of context-aware reranking strategy and MT model fine-tuned on the in-domain dataset is helpful to improve the performance. Because it can mitigate the problem that the inconsistency in transcripts caused by the lack of context. Finally, we use VITS model provided officially to reproduce audio files from the translation hypothesis.

1 Introduction

In this year, there is only one track in the speech to speech translation task which is the English to German translation (En-De) (Anastasopoulos et al., 2022). The audio files in English are given in the dataset, and we are required to produce the audio files in German. In recent research of speech to speech task, there are basically two paradigms with respect to the system architecture, which are cascade and end-to-end. And the cascade pipeline composed by an ASR model, a MT model and a TTS model is commonly used, because this system is more mature than end-to-end one. The advantage of this pipeline is that each module of the system can be a state-of-the-art one trained on sufficient independent corpora. It also allows us to perform experiments with different combinations of ASR models, MT models and TTS models. But compared to end-to-end system, this cascade system may not capture all information like accent of speakers, emotion, etc.

End-to-End system like S2UT is introduced in (Lee et al., 2021), which can be directly trained on speech to speech dataset with the help of text generation as the auxiliary task. However, we didn't adopt this approach due to the insufficiency of available corpora.

Dataset	Number of Utterance	Duration(hrs)
LibriSpeech	281,241	960.85
MuST-C	340,421	590.67
IWSLT	170,229	254.41
CoVoST	1362,422	1802.52
TEDLIUM3	268,214	453.42

Table 1: Data statistics of our ASR corpora

For the ASR model, we tried Conformer (Gulati et al., 2020), S2T Transformer (Synnaeve et al., 2019) and U2(Zhang et al., 2020), and obtained three types of ASR results.

In translation, inconsistency of translation of same words in the context is a common difficulty. This is caused by the flaw of conventional translation that treats each sentence independently in a documents, ignoring surrounding contexts. For example, a family name in English can be translated in different ways in Chinese. Because Chinese transcripts comes from transliteration, and there are lots of words share same pronunciation but different spelling. This may cause the ambiguity in transcripts, which is hard for readers to understand. To solve the problem, we propose the context-aware reranking strategy in translation, essentially an approach to adapt sentence-level MT models into document-level translation scenarios. It aims to generate the best candidate by taking previous contexts into account and reranking with scores estimated by all models.

2 Method

2.1 Data Preprocessing

We consider five datasets as our training set of ASR models, which are MuST-C V2 (Cattoni et al.,

Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 293 - 297

Language	WMT Bilingual	In-domain Text
En-De	79M	459K
En-Zh	96M	590K
En-Ja	42M	552K

Table 2: Data statistics of our MT corpora

2021), LibriSpeech(Panayotov et al., 2015), TED-LIUM 3 (Hernandez et al., 2018), CoVoST (Wang et al., 2020) and IWSLT. The statistical description is shown in Table 1. The CoVoST dataset has the longest duration and the largest number of utterances.

In the first step, we load the waveform of audio files as tensors and extract the 80-dimensional filter bank features of them. Because the encoder and decoder of a Transformer (Vaswani et al., 2017) model can only process limited size of sequences, we restrict the frame size of input speeches to the range of 50 to 3000, and the number of tokens should be no more than 150. At the same time, we calculate the speed of the speech by length of references and frame size of each sample. This metric could help us find those speech with small frame size but large number of tokens, or vice versa, which should be considered as outliers. So we choose the speech with the speed within $\mu(\tau) \pm 4 \times \sigma(\tau)$, where $\tau = \frac{\text{\# frames}}{\text{\# tokens}}$. Through these process pipeline in fine-grained level, we obtain the cleaned training set.

For the test set, we use the official dataset provided audios in the task. We also use the MuST-C dev, tst-COMMON and tst-HE set to evaluate our model so that they can be compared easily with other approaches.

For the training set of MT models, we follow the configuration and preprocessing procedures as (Wei et al., 2021), and the scale of the dataset is shown in Table 2.

2.2 Automatic Speech Recognition

We apply Conformer (Gulati et al., 2020) and S2T-Transformer (Synnaeve et al., 2019) to predict the fundamental results in an ensemble approach, and clean the predicted candidates with the U2 model (Zhang et al., 2020). All of these models are trained on the united dataset with the domain controlled training/generation (Wang et al., 2021). We ensemble the ASR result of the two models, and some results have been corrected in the post-processing. Sometimes both Conformer and S2T-Transformer makes errors in the recognising process, except the errors appeared in different position. For example, in a same sentence, the Conformer would recognise the "ex-boyfriend" as "next boyfriend" incorrectly, and the S2T-Transformer may misidentify "the cuss words" as "the cusp words". Through ensembling, these errors can be eliminated and results can be improved. We proved that the ensembling of these heterogeneous ASR models can in some what extent improve the possibility of choosing the correct answer.

Algorithm 1 Context-aware Translation reranking

Require: MT, MT', LM, context length, beam size, utterance list: $\mathcal{F}, \mathcal{G}, \mathcal{Q}, N, k, S$
  Initialize: Context Buffer $C \leftarrow \{\}$
  Initialize: source text index $i \leftarrow 0$
  while $i \neq |S| - 1$ **do**
    $\hat{Y}, P_{\text{f}} \leftarrow \mathcal{F}(u_i, k)$: propose candidates
    $P_{\text{g}} \leftarrow \mathcal{G}(u_i, \hat{Y})$: scoring with MT'
    if $i < N$ **then**
      $P_{\text{q}} \leftarrow \mathcal{Q}(\hat{Y}, C)$
    else
      $P_{\text{q}} \leftarrow \mathcal{Q}(\hat{Y}, C_{[-N:]})$
    end if
    $\hat{y}^* \leftarrow \arg\max_{\hat{y}} \sum m \in \{f, g, q\} w_m \log P_m$
    $C \leftarrow C \cup \{\hat{y}^*\}$
    $i \leftarrow i + 1$
  end while
  return C

Meanwhile, we find that two autoregressive models both have the drawback of producing meaningless sentences when acoustic inputs are applause or laughing from the audience. In this situation, U2 presents the stability and robustness in predicting those audio without real utterances. So, we use U2 as the criteria to filter the ensemble results comes from Conformer and S2T-Transformer. It means, for each sample, we predict with U2 first and see if the prediction is a blank line, if it is, we directly use it as the output, otherwise, we predict the sample again with the ensembled model mentioned above. This is the key to apply U2, but it would not change any other prediction of ensemble results.

After the cleaning process of U2, results are more anti-interference to the sample that filled with laughter or meaningless natural noise.

Test set	Approach	BLEU	ChrF	TER	Perf. Drop
dev	Oracle	32.1	0.61	0.534	21.4%
	TTS	25.12 (-6.98)	0.58 (-0.03)	0.585 (+0.051)	
tst-HE	Oracle	34.0	0.63	0.498	28.82%
	TTS	24.2 (-9.8)	0.56 (-0.07)	0.609 (+0.111)	
tst-COMMON	Oracle	31.2	0.63	0.550	21.80%
	TTS	24.4 (-6.8)	0.57 (-0.06)	0.627 (+0.077)	

Table 3: This table presents our overall performance evaluated on MuST-C dev, tst-HE and tst-COMMON set. Oracle stands for directly evaluating translation outputs of the MT model. TTS stands for evaluating on the transcripts predicted from the TTS output. Note that all results are evaluated without punctuation and with lower-casing since the wav2vec ASR model is only able to predict in that form. The column "Perf. Drop" statistics the drop of BLEU when applied with TTS.

2.3 Translation Models

We use the WMT21 news corpora to train the MT model in En-De direction, then, use the combination of MuST-C and IWSLT dataset to fine-tune the pretrained model.

2.4 Context-aware MT reranking

Following the work in (Yu et al., 2020) that utilises the noisy channel model (Brown et al., 1993) in document-level translation, we adopted similar strategy to improve the translation with longer context information. However, we make some simplification on the decoding process and the scoring function. More specifically, we restrict the context to a sliding window that only taking a fixed size of sentences into account when applying the LM scoring:

$$\begin{aligned}\mathcal{O}(x, y^{-N:}, y^i) =& w_{\text{MT}} \log p_{\text{MT}}(y^i|x^i) \\ &+ w_{\text{LM}} \log p_{\text{LM}}(y^i|y^{-N:}) \\ &+ w_{\text{MT}'} \log p_{\text{MT}'}(x^i|y^i) \quad (1)\end{aligned}$$

where N is the context length, w are weights for each component. The decoding process is also simplified into a greedy search instead of sentence-level beam search as described in Algo 1. During inference, we find that the test set is exactly same as the tst2022-en-de used in the offline, therefore, we manually regroup ASR outputs back to documents and translate them with this approach.

2.5 Text to Speech

In a cascade speech to speech translation system, text to speech (TTS) is the final module to convert translations into speech. We use the pretrained VITS (Kim et al., 2021) model for this procedure. VITS adapts variatoinal inference augmented with normalizing flows and an adversarial training process, largely improving the quality of generated speech. During inference process we only need to provide German texts, and use the model to produce raw audio files with 22kHz sample rate.

3 Experiments

3.1 Setup

In the training of our ASR models, we use the sentencepiece model (Kudo and Richardson, 2018) for tokenization with vocab size=20000. Configurations of ASR models are exactly same to our offline submission. We follow the recipe of (Wei et al., 2021) to train our NMT models in both directions, as well as the language model. All MT models are also fine-tuned on in-domain corpora for additional 10 epochs. We implemented all models with `fairseq` (Ott et al., 2019).

The automatic evaluation of our S2S system is achieved by calculating metrics on the re-transcribed outputs from our system. Specifically, an officially assigned ASR model: "wav2vec2-large-xlsr-53-german" (Baevski et al., 2020) is used to transcribe the TTS generated audio files back to texts first. Then, they are used for the evaluation with automatic tools performed in text-level. This significantly reduces the difficulty of evaluation but still preserves the fairness. We use BLEU (Papineni et al., 2002), ChrF (Popovic, 2015) and TER (Snover et al., 2006) as evaluation metrics in our experiments.

3.2 Results

Because the speech cannot be directly compared to transcripts, we have to convert the speech into transcripts by the Wav2vec ASR model. We tested

ASR Model	CoVoST	MuST-C	TEDLIUM3	LibriSpeech
w/ Domain Tag	11.27	6.31	5.33	4.39
wo/ Domain Tag	17.56	15.58	8.72	7.98

Table 4: Comparison of wer scores of ASR model trained on dataset with domain tag or not.

the score of BLEU, ChrF and TER by evaluating the translation outputs of MT model and the re-transcribed results of final outputs from TTS. And those scores can be seen in Table 3. Note that before computing evaluation metrics, we applied some normalizing process to make the results of Oracle and TTS comparable. More specifically, since the re-transcribed text from the wav2vec model is lower-cased and has no punctuation, we also perform lower-casing and removing of punctuation for Oracle hypothesis and the references. Finally, we evaluate metrics on Oracle and TTS hypothesis towards the normalized references.

From the experimental results on three sub sets of MuST-C, we have some interesting findings. Through the process of TTS and re-ASR, the BLEU score and ChrF score has both decreased by about 7+ and 0.05+, and the TER score increased by 0.07+. This trend appears in both three test sets, demonstrating that there might be serious information loss in this process. However, further conclusions can only be drawn from the human evaluation.

3.3 Ablation

Effectiveness of domain controlled generation We test whether the domain tag prefix is useful for the performance of model, and the results are shown in Table 4. There are four domain tag used in our new dataset, including "<MC>", "<LS>", "<TL>" and "<CV>". All these prefix represents the abbreviations of each dataset. Compared with the results of model fed by dataset without using any domain prefix tags, the model trained on the tagged dataset has the better performance. This essentially benefits from the extra prior information provided by the domain prefix tags. In detail, domain tags provides more latent information that cannot be easily captured in raw audios, making the generation more deterministic. Meanwhile, this allows us to control the generation style in our demanded domain, being closer to the reference. So, the domain tag prefix effectively improves the performance of our model.

4 Conclusion

In the paper, we elaborate the cascade system for this Speech to Speech task. There are several strategies we applied to improve the system, including domain-tag prefix and the context-aware reranking strategy. We did some experiments to verify the reliability of those strategies for a cascade system, and we also made some analysis from the theoretical level. In the future, we are going to explore the feasibility of the end-to-end system, since it might reduce the negative impact of information loss on system performance.

References

Antonios Anastasopoulos, Luisa Bentivogli, Marcely Z. Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Marcello Federico, Christian Federmann, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel M. Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Juan Pino, Elizabeth Salesky, Jiatong Shi, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alex Waibel, Changhan Wang, and Shinji Watanabe. 2022. FINDINGS OF THE IWSLT 2022 EVALUATION CAMPAIGN. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, Dublin, Ireland. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311.

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. *Comput. Speech Lang.*, 66:101155.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang,

Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 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*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Miguel Pino, and Wei-Ning Hsu. 2021. Direct speech-to-speech translation with discrete units. *CoRR*, abs/2107.05604.

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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

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 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus.

Minghan Wang, Yuxia Wang, Chang Su, Jiaxin Guo, Yingtao Zhang, Yujia Liu, Min Zhang, Shimin Tao, Xingshan Zeng, Liangyou Li, Hao Yang, and Ying Qin. 2021. The hw-tsc's offline speech translation systems for IWSLT 2021 evaluation. *CoRR*, abs/2108.03845.

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, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 225–231. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Trans. Assoc. Comput. Linguistics*, 8:346–360.

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. *CoRR*, abs/2012.05481.