HW-TSC's Speech to Text Translation System for IWSLT 2024 in Indic track

Bin Wei, Zongyao Li, Jiaxin Guo, Daimeng Wei, Zhanglin Wu, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{weibin29,lizongyao, guojiaxin1, weidaimeng, wuzhanglin2, chenxiaoyu35, raozhiqiang, lishaojun18, luoyuanchang1, shanghengchao,yanghao30,jiangyanfei}@huawei.com

Abstract

This article introduces the process of HW-TSC and the results of IWSLT 2024 Indic Track Speech to Text Translation. We designed a cascade system consisting of an ASR model and a machine translation model to translate speech from one language to another. For the ASR part, we directly use whisper large v3 as our ASR model. Our main task is to optimize the machine translation model (en2ta, en2hi, en2bn). In the process of optimizing the translation model, we first use bilingual corpus to train the baseline model. Then we use monolingual data to construct pseudo-corpus data to further enhance the baseline model. Finally, we filter the parallel corpus data through the labse(Feng et al., 2022) filtering method and finetune the model again, which can further improve the BLEU score. We also selected domain data from bilingual corpus to finetune previous model to achieve the best results.

1 Introduction

This article describes the Indic track speech-to-text translation task submitted by HW-TSC at IWSLT 2024.

From a system architecture perspective, current research on speech-to-text translation can be divided into two forms: end-to-end and cascade systems. Cascade systems usually consist of a speech recognition (ASR) module and a text-to-text machine translation (MT) module. Although integrating these modules may be complex, the results are still very satisfactory as long as there are sufficient data resources to train each module. Additionally, the end-to-end approach can generate translation results directly from the unified model with speech input. However, what we need to know is that the parallel data required to train an end-to-end speech translation model is extremely scarce.

2 Methods

Our approach ultimately adopts a cascade approach.

2.1 ASR

In our cascaded system we have whisper-large-v3 as our ASR module. The researchers of Whisper(Radford et al., 2023) has scaled up the supervised speech recognition dataset from thousands to 680,000 hours. Pretraining on such a large-scale weakly supervised dataset enables the model to be applicable to various data types or domains. Furthermore, Whisper has expanded the scope of weakly supervised pretraining to include multilingual and multitask scenarios. Therefore, we ultimately chose the powerful recognition-capable Whisper-large-v3 model as our ASR module.

2.2 MT

Our cascade system includes the Transformer (Vaswani et al., 2017) as the MT module, which has become a prevalent method for machine translation in recent years. The Transformer has achieved impressive results, even with a primitive architecture that requires minimal modification. To improve the offline MT model performance, we utilize multiple training strategies.

2.2.1 labse

Language-agnostic BERT Sentence Embedding (Feng et al., 2022) is an effective parallel corpus filtering method, which can effectively filter out high-quality bilingual data. We can use the filtered high-quality bilinguals and then finetune our model. Finally, we applied this method to this competition, which greatly improved the results in the three directions. In this experiment, we get 37 million filtered high-quality bilinguals in the en2ta direction, 55 million filtered high-quality bilinguals in the en2hi direction, and 43 million filtered high-quality

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 101–104

bilinguals in the en2bn direction from bilingual data.

2.2.2 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a simple but effective strategy to boost neural machine translation (NMT) (Bahdanau et al., 2015) performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. This method is more effective than knowledge distillation and dual learning. Finally,

2.2.3 Forward Translation

Forward translation (FT) (Abdulmumin, 2021) uses source-side monolingual data to improve model performance. The general procedure of FT involves three steps: (1) randomly sampling a subset from large-scale source monolingual data; (2) using a "teacher" NMT model to translate the subset into the target language, thereby constructing synthetic parallel data; and (3) combining the synthetic and authentic parallel data to train a "student" NMT model.

2.2.4 Back Translation

Augmenting parallel training data with back-translation (BT) (Sennrich et al., 2016; Wei et al., 2023) has been shown effective for improving NMT using target monolingual data. Numerous works have expanded the understanding of BT and investigated various approaches to generate synthetic source sentences. Edunov et al. found that back-translations obtained via sampling or noised beam outputs tend to be more effective than those via beam or greedy search in most scenarios. For optimal joint use with FT, we employ sampling back-translation (ST) (Edunov et al., 2018).

2.2.5 Domain Fine-tuning

Previous studies have shown that fine-tuning a model with in-domain data can significantly enhance its performance. We use the model scoring method to select data from the bilingual training data that are close to the dev set in domain, and then use these domain data to finetune the model, which can further improve the result. Finally, we select 12 million domain data in the en2ta direction, 15 million domain data in the en2hi direction, and 10 million domain data in the en2bn direction from the bilingual training data.

2.2.6 Regularized Dropout

Regularized Dropout (R-Drop) (Wu et al., 2021) improves performance over standard dropout, especially for recurrent neural networks on tasks with long input sequences. It ensures more consistent regularization while maintaining model uncertainty estimates. The consistent masking also improves training efficiency compared to standard dropout. Overall, Regularized Dropout is an enhanced dropout technique that often outperforms standard dropout.

3 Experiments Setup

3.1 ASR

In our cascade system, we use whisper-large-v3 as our ASR module, which we will not introduce here.

3.2 MT

3.2.1 Model

For our experiments using the MT model, we utilize the Transformer deep model architecture. The configuration of the MT model is as follows:n_encoder layers = 35, n_decoder layers = 3, n_heads = 8, d_hidden = 512, d_FFN = 2048.

3.2.2 Dataset

To train the MT model, we collected all available parallel corpora from the official website and selected paralla data similar to the dev domain. The amount of data is shown in Table 1. We first trained respective baseline models in the three directions using bilingual data. Then, we construct pseudo-corpus based on existing monolingual data in each language direction to gradually enhance the baseline model.

	Bilingual	Source	Target
en-ta	57M	200M	70M
en-hi	80M	200M	230M
en-bn	82M	200M	190M

Table 1: Bilingual and monolingual data used for training.

3.2.3 Training

We utilize the open-source Fairseq (Ott et al., 2019) for training, with the following main parameters: each model is trained using 8 GPUs, with a batch size of 2048, a parameter update frequency of 32,

and a learning rate of 5e-4. Additionally, a label smoothing value of 0.1 was used, with 4000 warmup steps and a dropout of 0.1. The Adam optimizer is also employed, with ß1 = 0.9 and ß2 = 0.98. During the inference phase, a beam size of 4 is used. The length penalties are set to 1.0.

3.3 Results

We can see results From Table 2, In the field of machine translation, Domain Finetuning, Forward Translation, and labse filter method are frequently employed methods to enhance translation quality. It is evident from Table 4 that these training strategies can effectively improve the overall quality of the system.

Language-pair	Training strategies	Bleu
en-hi	Bilingual baseline	51.9
	+ FT+BT	53.8
	+ labse Bilingual Finetune	54.7
	+ Domain Finetune	64.8
en-ta	Bilingual baseline	41.9
	+ FT+BT	42.2
	+ labse Bilingual Finetune	43.1
	+ Domain Finetune	45.2
en-bn	Bilingual baseline	38
	+ FT+BT	40.4
	+ labse Bilingual Finetune	42.1
	+ Domain Finetune	44.8

Table 2: All the results for dev testsets in three directions(EN-HI,EN-TA,EN-BN).FT means Forward Translation. BT means Back Translation.

At the same time, we also calculated the blue of NLLB-200-3.3B (Costa-jussà et al., 2022) in three directions, as shown in Table 3, for comparison with our results. As can be seen from Table 2 and Table 3, our model is far better than the NLLB model.

Language-pair	NLLB baseline
en-hi	40.9
en-ta	20.4
en-bn	25.7

Table 3: NLLB-200-3.3B results for dev testsets in three directions(EN-HI,EN-TA,EN-BN).

4 Conclusion

In this paper, we report on our work on IWSLT2024 speech-to-text translation evaluation in Indic Track. We mainly introduce our cascade system and the main optimization processes and methods of the MT model. We improve the final results by focusing on optimizing the MT model. For cascade systems, the impact of the MT model on the results is crucial. For the future we plan to further explore the direction of end-to-end systems.

References

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

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

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

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

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

pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

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.

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

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