HW-TSC's Submissions To the IWSLT2024 Low-resource Speech Translation Tasks

Jiawei Zheng, Hengchao Shang, Zongyao Li, Zhanglin Wu, Daimeng Wei, Zhiqiang Rao,
Shaojun Li,Jiaxin Guo,Bin wei,Yuhao Xie,Yuanchang Luo, Hao Yang
Huawei Translation Service Center, Beijing, China
{zhengjiawei15,shanghengchao,lizongyao,wuzhanglin2, weidaimeng, raozhiqiang,lishaojun18,
guojiaxin1,weibin29,xieyuhao2,luoyuanchang1,yanghao30}@huawei.com

Abstract

In this work, we submitted our systems to the low-resource track of the IWSLT 2024 Speech Translation Campaign. Our systems tackled the unconstrained condition of the Dialectal Arabic North Levantine (ISO-3 code: apc) to English language pair. We proposed a cascaded solution consisting of an automatic speech recognition (ASR) model and a machine translation (MT) model. It was noted that the ASR model employed the pre-trained Whisper-large-v3 model to process the speech data, while the MT model adopted the Transformer architecture. To improve the quality of the MT model, it was stated that our system utilized not only the data provided by the competition but also an additional 54 million parallel sentences. Ultimately, we reported that our final system achieved a BLEU score of 24.7 for apc-to-English translation.

1 Introduction

The IWSLT 2024 Speech Translation Campaign featured a low-resource track that posed a challenging task: translating dialectal Arabic speech to English text. This language pair is particularly demanding due to the scarcity of available training data and the complexity of handling dialectal Arabic variations. To tackle this problem, we propose a cascaded approach that leverages state-of-the-art models for automatic speech recognition (ASR) and machine translation (MT).

End-to-end models, which directly map audio inputs to translated text outputs, heavily rely on the availability of paired audio and transcription data. For low-resource tasks, acquiring such data can be exceptionally challenging. Consequently, we adopted a cascaded model architecture, which decouples the speech recognition and translation components. This approach allows us to leverage additional parallel text data to enhance the MT module's performance, ultimately benefiting the overall speech translation task.

Our ASR model is built upon the whisper-large-v3 architecture, a powerful pre-trained model that has demonstrated impressive performance in transcribing diverse speech data.By employing this model, we aim to accurately transcribe the dialectal Arabic speech inputs, to capture the nuances and variations present in the spoken language.

For the MT component, we adopt the Transformer architecture(Vaswani et al., 2017), which has become the de facto standard for modern neural machine translation systems. The Transformer model is known for its ability to effectively capture long-range dependencies and produce high-quality translations, making it well-suited for the task at hand.

To further enhance the performance of our MT system, we augment the provided training data with a substantial amount of additional parallel data, totaling 54 million sentence pairs. This data augmentation strategy aims to improve the model's robustness and generalization capabilities, enabling it to better handle the complexities of translating between dialectal Arabic and English.

By combining the strengths of these two powerful models in a cascaded fashion, we aim to deliver a robust and accurate speech-to-text translation system for the IWSLT 2024 low-resource track, pushing the boundaries of what is achievable in this challenging language pair.

2 Data

2.1 Data Source

We used two sets of data to train our machine translation (MT) model. Firstly, we utilized a dataset provided by the IWSLT 2024 competition, comprising approximately 42 million lines of MSA-English bilingual data. This data is sourced from various platforms including Opensubtitles[1], UN[2],QED[3]

[1] https://opus.nlpl.eu/OpenSubtitles-v2018.php
[2] https://conferences.unite.un.org/UNCorpus
[3] https://opus.nlpl.eu/QED-v2.0a.php

Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024), pages 208–211
August 15-16, 2024 ©2024 Association for Computational Linguistics

,TED[4],GlobalVoices[5],News-Commentary[6]. These datasets are of high quality and provide a solid foundation for our MT model to learn the correspondences and translation patterns between the two languages.

However, due to the inherent differences between dialectal Arabic and MSA, relying solely on the official dataset may not cover all linguistic phenomena of the target language pair. Therefore, to enhance the generalization capability of our MT model, we additionally utilized approximately 73 million lines of Arabic-English bilingual data. These datasets cover a wider range of dialectal language phenomena, enabling our MT model to better understand dialectal Arabic and produce more accurate and fluent translations into English. The data size is shown in Table 1.

Data Source	Volume
IWSLT 2024 Official Dataset	42M
Additional Arabic-English Bilingual Data	73M

Table 1: Uncleaned Bilingual used for training.

2.2 Data Pre-processing

The data preprocessing pipeline follows our previous work (Wei et al., 2021). We employed various strategies, including deduplication, XML content processing, language identification based on langid (Lui and Baldwin, 2012), and filtering using fast-align (Dyer et al., 2013). These preprocessing steps help improve the quality of the corpus and ensure consistency in the training data.

For the sake of conciseness, we will not elaborate on the specific details of the preprocessing steps. Interested readers can refer to the relevant papers for more information. Overall, this established set of data preprocessing strategies provides high-quality training data for our machine translation system, laying the foundation for achieving excellent translation performance. The size of the preprocessed data is shown in Table 2.

3 Methods

We employed a cascade approach for the Spoken Language Translation (ST) task, leveraging both Automatic Speech Recognition (ASR) and Machine Translation (MT) models. The cascade model consists of two stages: the ASR stage and the MT stage.

Data Source	Volume
IWSLT 2024 Official Dataset	27M
Additional Arabic-English Bilingual Data	54M
Total	81M

Table 2: Cleaned Bilingual used for training.

3.1 ASR

The automatic speech recognition (ASR) module plays a crucial role in speech-to-text translation systems. To obtain high-quality speech transcriptions, we chose to employ the whisper-large-v3 model proposed by OpenAI as our system's ASR module.

Whisper(Radford et al., 2023) is a powerful speech recognition model that has learned to map raw audio to speech units through self-supervised pretraining. It not only excels in high-resource languages like English but also demonstrates outstanding performance in various low-resource languages. The latest large-v3 version further scales up the model size and leverages larger datasets for pretraining, thereby enhancing its recognition accuracy.

One of the primary motivations for adopting the whisper-large-v3 model, is its robust ability to handle diverse language variations and accents. Dialectal Arabic exhibits a rich variety of speech variations, necessitating the ASR system to possess sufficient robustness to accommodate these differences. Whisper, with its powerful modeling capabilities, can effectively adapt to such speech diversity, laying the foundation for subsequent machine translation processes.

3.2 MT

Our cascaded system utilizes the Transformer architecture as the MT module, which has become the predominant approach for machine translation in recent years. Remarkably, the Transformer achieves impressive results even with its original architecture requiring minimal modifications. To further boost the performance of our offline MT model, we employed a variety of training strategies.

3.2.1 LaBSE

LaBSE (Feng et al., 2020) acts as a natural filter for parallel corpora, efficiently extracting high-quality bilingual data. We can utilize this filtered

[4] https://opus.nlpl.eu/TED2020-v1.php

[5] https://opus.nlpl.eu/GlobalVoices-v2017q3.php

[6] https://opus.nlpl.eu/News-Commentary-v16.php

high-quality bilingual data to fine-tune our models, thereby acting as a natural denoising process. We applied this method to the current competition, resulting in a subtle improvement in BLEU scores.

3.2.2 Curriculum Learning

A practical curriculum learning (CL) approach for NMT should address two key issues: ranking training examples by difficulty and modifying the sampling procedure based on ranking (Zhang et al., 2019). For ranking, we estimate example difficulty using domain features (Wang et al., 2020). The domain feature is calculated as:

$$q(x, y) = \frac{\log P(y|x;\theta_{in}) - \log P(y|x;\theta_{out})}{|y|} \tag{1}$$

Where θ_{in} is an in-domain NMT model, while θ_{out} is an out-of-domain model. The novel domain is treated as in-domain.

We fine-tune the model on the valid set to get the teacher model and select top 40% of the highest scoring data for finetuning.

3.2.3 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.

4 Experiments

4.1 ASR

Whisper is a powerful speech recognition model based on self-supervised pretraining, exhibiting exceptional performance across multiple languages, particularly in handling diverse speech variations and accents. Given the rich variety of dialects in Arabic, the robustness of the ASR model is paramount. With its outstanding modeling capabilities, Whisper can effectively adapt to such speech diversity. Considering Whisper's remarkable performance in multilingual ASR tasks, we directly employed its latest large-v3 version without any modifications to the model itself. Our objective is to fully leverage this powerful pretrained model as the core ASR component within our cascaded system, providing high-quality speech transcription inputs for the entire speech translation pipeline.

4.2 MT

Model For our experiments using the MT model,we utilize the Transformer deep model architecture.The configuration of the MT model is as follows:nencoder layers = 35, ndecoder layers = 3, nheads =8, dhidden = 512, dF F N = 2048.

Training We use SacreBLEU (Post, 2018) to measure system performances. 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 4, 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 1 = 0.9 and 2 = 0.98. During the inference phase, a beam size of 4 is used. The length penalties are set to 1.0.

5 Results

The multi-step fine-tuning method first pretrains a base model on large-scale general-domain corpora, and then conducts multiple rounds of fine-tuning on the task-specific data, with each round optimizing the model based on the previous round. This approach leverages general knowledge, addresses data distribution mismatch issues, and avoids overfitting. After each round of fine-tuning, the BLEU metric is used to evaluate the translation quality, serving as the basis for determining whether to proceed with the next round of fine-tuning. Through this gradual fine-tuning process, the model's performance can be progressively enhanced.Table 3 shows our baseline results and the fine-tuning results at each step.

Traing Stategies	BLEU
All Bilingual baseline	17.6
+ LaBSE bitext Finetune	17.7
+ Curriculum Learning +R-Drop	24.7

Table 3: BLEU scores of apc→en NMT system on IWSLT low-resource test set.

5.1 Ablation study of different bilingual data

According to the experimental results shown in Table 4, we conducted an ablation study to determine whether the additional bilingual data contributes to improving the performance of the machine translation model. We can clearly see that by adding Arabic-to-English bilingual data, the model can better capture dialectal Arabic knowledge.

Traing Stategies	BLEU
IWSLT 2024 Official Bilingual baseline	15.7
All Bilingual baseline	17.6

Table 4: BLEU Scores for Different Bilingual Data

6 Conclusion

Our research has led to the following key conclusions: Firstly, for the Arabic-to-English translation task, incorporating additional bilingual corpus data significantly enhanced model performance. These corpora contained rich knowledge of Arabic dialects, enabling the model to better learn and translate these special language variants, thereby improving the overall translation quality. Secondly, we adopted advanced training strategies such as Curriculum Learning and R-drop, which also brought substantial performance gains to the machine translation model. Curriculum Learning facilitated gradual learning from easy to difficult scenarios, while R-drop effectively mitigated overfitting issues and improved the model's generalization capability. These strategies were the core methods employed in our submission, yielding outstanding practical results.

References

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

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

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. 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.

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.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 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.

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.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.