The USTC's Dialect Speech Translation System for IWSLT 2023

Pan Deng[1] **Shihao Chen**[1] **Weitai Zhang**[1,2] **Jie Zhang**[1] **Lirong Dai**[1]
[1]University of Science and Technology of China, Hefei, China
[2]iFlytek Research, Hefei, China
{pdeng, shchen16, zwt2021}@mail.ustc.edu.cn; {jzhang6, lrdai}@ustc.edu.cn

Abstract

This paper presents the USTC system for the IWSLT 2023 Dialectal and Low-resource shared task, which involves translation from Tunisian Arabic to English. We aim to investigate the mutual transfer between Tunisian Arabic and Modern Standard Arabic (MSA) to enhance the performance of speech translation (ST) by following standard pre-training and fine-tuning pipelines. We synthesize a substantial amount of pseudo Tunisian-English paired data using a multi-step pre-training approach. Integrating a Tunisian-MSA translation module into the end-to-end ST model enables the transfer from Tunisian to MSA and facilitates linguistic normalization of the dialect. To increase the robustness of the ST system, we optimize the model's ability to adapt to ASR errors and propose a model ensemble method. Results indicate that applying the dialect transfer method can increase the BLEU score of dialectal ST. It is shown that the optimal system ensembles both cascaded and end-to-end ST models, achieving BLEU improvements of 2.4 and 2.8 in test1 and test2 sets, respectively, compared to the best published system.

1 Introduction

In this paper, we present the USTC's submission to the Dialectal and Low-resource track of IWSLT 2023 Evaluation Campaign (Agarwal et al., 2023), aiming to translate Tunisian Arabic speech to English text. Modern Standard Arabic (MSA) is the official language of Arabic-spoken countries. However, Arabic dialects like Tunisian and Egyptian are prevalent in everyday communication, exhibiting a similar relation between Chinese and Cantonese. MSA benefits from an abundant supply of unlabeled speech and text data, as well as relatively adequate automatic speech recognition (ASR) and machine translation (MT) paired data. In contrast, dialectical forms of Arabic have much less paired data and more irregularities in both pronunciation and writing (Ben Abdallah et al., 2020).

This paper aims to explore the transfer between high-resource MSA and low-resource Tunisian dialects, as well as effective training and decoding strategies for speech translation (ST) tasks related to low-resource dialects. To facilitate **dialect transfer**, we introduce two approaches. Firstly, we pre-train a model using high-resource MSA data, which is then fine-tuned using low-resource Tunisian data. This approach involves transferring model parameters and can be used to train various models, e.g., ASR, MT, end-to-end ST. Secondly, we also develop two transformation models for explicit dialect transfer. On one hand, for the augmentation of MT data, we build an MT model that translates MSA into Tunisian, resulting in a vast amount of pseudo Tunisian-English paired data. On the other hand, the Tunisian-MSA MT encoder module is built and then integrated into the end-to-end ST model, which can implicitly normalize dialectal expressions. In addition, we also propose robust training and decoding strategies from two perspectives. To improve the robustness of the MT model against ASR errors, we fine-tune the MT model with the ASR output from the CTC (Graves et al., 2006) layer or the ASR decoder. The model ensemble method is exploited to decode multiple models synchronously, which is shown to be rather beneficial for the performance.

The rest of this paper is organized as follows. Section 2 describes data preparation (e.g., datasets, pre-processing). Section 3 presents the methods for training and decoding ASR, MT and ST models. Experimental setup and results are given in Section 4. Finally, Section 5 concludes this work.

2 Data Preparation

2.1 Datasets

In this year's shared task, there are two types of data conditions: constrained and unconstrained. In order to provide a fair comparison with last year's

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 102–112

Task	Dataset	Condition	Utterances	Hours
ASR	Tunisian	A	0.2M	160
	MGB2	B	1.1M	1100
	MGB2+Private data	C	3.4M	4600
ST	Tunisian	A	0.2M	160

Table 1: The summary of the Audio data.

	Dataset	Condition	Ta-En	MSA-En
Collected	Tunisian	A	0.2M	-
	OPUS	B	-	42M
	OPUS+Private data	C	-	61M
Filtered	Tunisian	A	0.2M	-
	OPUS	B	-	32M
	OPUS+Private data	C	-	47M

Table 2: The summary of the text data.

Translation direction	Training data	MT model
Tunisian-English English-Tunisian	Ta-En En-Ta	Ta2En En2Ta
MSA-English English-MSA	MSA-En En-MSA	MSA2En En2MSA
Tunisian-MSA MSA-Tunisian	Ta-MSA MSA-Ta	Ta2MSA MSA2Ta
Tunisian-MSA-English	Ta-MSA-En	-

Table 3: Summary of abbreviations used in this paper.

results, we subdivided the unconstrained condition into the dialect adaption condition and the fully unconstrained condition. For convenience, we denote the constrained condition as **condition A**, the dialect adaption condition as **condition B**, and the fully unconstrained condition as **condition C**.

Table 1 summarizes statistics of the ASR and ST datasets. The Tunisian dataset[1] in condition A is Arabic dialect data. In addition to the MGB2 data (Ali et al., 2016) of condition B, we used additional private data mainly from MSA for ASR training in condition C. Table 2 summarizes the statistics of the MT datasets. The MT data for condition A are Tunisian-English (Ta-En) paired data, while for condition B/C, the MT data consist of MSA-English (MSA-En) paired data(Tiedemann and Thottingal, 2020). All MT data undergoes preprocessing, which includes cleaning and filtering. Table 3 summarizes the abbreviations for MT models and training data associated with the translation direction that are used in the sequel.

[1]The LDC Catalog ID of the Tunisian dataset for IWSLT is LDC2022E01.

2.2 Audio data pre-processing

As the audio data of condition B/C had a sampling rate of 16kHz, we upsampled the speech signal in the Tunisian dataset from 8kHz to 16kHz using the sox toolkit[2]. We extracted 40-dimensional log-mel filterbank features with a frame length of 25ms and a frame shift of 10ms, and then normalized these features with a zero mean and unit variance. We applied SpecAugment (Park et al., 2019) in the time dimension with mask parameters (m_T, T) = (2, 70). Afterwards, we filtered out audio data that is longer than 3k frames. Further, we introduced speech perturbations at ratios of 0.9 and 1.1.

2.3 Text Processing & Filtering

We kept the MSA and Tunisian text data in their original form without any normalization such as removing diacritical marks or converting Alif/Ya/Ta-Marbuta symbols. We removed punctuations from MSA, Tunisian, and English text while we converted the English text to lowercase. Our data filtering process in condition B/C includes **Length Match** and **Inference Score**.

- **Length Match:** Text samples exceeding 250 words were dropped first. Next, we calculated the length ratio between the source and target language text. Text samples with length ratios exceeding 2 or below 0.4 were deemed to be length mismatching cases and were subsequently removed. As such, approximately 6M text data in condition B were eliminated.
- **Inference Score:** Initially, a basic MT model (scoring model) was trained on raw MSA-En data in condition B. Subsequently, the scoring model was used to infer the same MSA-En raw data, resulting in inference scores based on logarithmic posterior probabilities. Finally, MSA-En data associated with lower inference scores were removed, leading to another 4M text data being eliminated from condition B.

Table 2 summarizes the filtered data used for training. In total, 10M text data in condition B and 4M text data in condition C were removed.

3 Methods

3.1 Automatic Speech Recognition

We employed several ASR models with different structures in experiments, including the VGG-

[2]http://sox.sourceforge.net

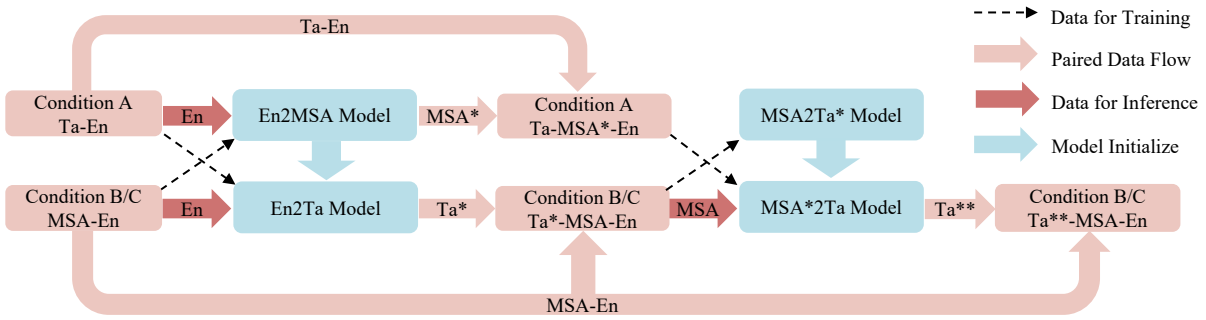

Figure 1: The data augmentation method for Tunisian-English Text, where * indicates the pseudo text.

Conformer model (Simonyan and Zisserman, 2014; Gulati et al., 2020), VGG-Transformer model (Vaswani et al., 2017) and GateCNN-Conformer model (Dauphin et al., 2017). These ASR models differ in their feature extractor modules (VGG, GateCNN) and acoustic modules (Conformer, Transformer). We chose diverse models with the expectation that increasing the variability of ASR models would improve the final ASR performance when using model ensemble methods. For dialect transfer in condition B/C, we pre-trained an ASR model using MSA data, which was then fine-tuned using the Tunisian data. Note that for condition A, we initially attempted to pre-train a phoneme recognition model for Tunisian but found it to be useless after fine-tuning the pre-trained model.

3.2 Data Augmentation for MT

We considered various data augmentation techniques for MT. To augment the Tunisian-English (Ta-En) dialect MT data, we used the back translation and forward translation (**BTFT**) method to create a synthetic parallel corpus that can be merged with the true bilingual data. To accomplish **dialect transfer** from MSA to Tunisian, we constructed a pivot MT model that converts MSA to Tunisian and produces abundant synthetic Ta-En data.

BTFT: Two MT models were first trained from Tunisian to English (Ta2En) and from English to Tunisian (En2Ta) using MT data of condition A. The Tunisian text and English text were then respectively fed to the corresponding MT models for inference, resulting in paired Tunisian to synthetic-English text and paired synthetic-Tunisian to English text. It is worth noting that the Ta2En model implements the forward translation approach similarly to the sequence-level knowledge distillation method (Kim and Rush, 2016), while the En2Ta model employs the backward translation (Sennrich et al., 2016a) approach. Ultimately, the obtained synthetic data and the original data were merged to form the BTFT dataset.

Dialect Transfer: In the IWSLT 2022 dialect ST track, (Yang et al., 2022) presented an effective Ta2En-bt-tune model that generates synthetic Tunisian-English data by converting MSA to pseudo-Tunisian with an MSA2Ta MT model. In Figure 1, we modified this approach by introducing a multi-step pre-training technique that improves the quality of pseudo-Tunisian and enhances downstream translation tasks. Our dialect transfer method is outlined as follows:

(1) Firstly, the En2MSA (English to MSA) model was pre-trained using condition B/C MT data and then fine-tuned using the MT data from condition A to create the En2Ta model.

(2) The En2MSA and En2Ta models were utilized separately with the English texts from condition A and condition B/C as inputs to generate paired Ta-MSA-En triple text data for condition A/B/C. The pseudo-text in condition A is the MSA* text, whereas the pseudo-text in condition B/C is the Tunisian* text (* representing pseudo-text). Notably, during this step, the pseudo-Tunisian* text derived from condition B/C is marked as the first iteration.

(3) Next, we trained an MSA2Ta (MSA to Tunisian) model, which serves as a pivot MT model. We pre-trained the model with the MSA-Ta* data of condition B/C and fine-tuned it using the MSA*-Ta data of condition A from step 2.

(4) Lastly, we input the MSA text of condition B/C to the MSA2Ta model for inference, generating the second iteration of the pseudo-Tunisian text (marked as pseudo-Tunisian**). We re-created the paired triple text data of Ta-MSA-En text by merging the pseudo-Tunisian** text with the primary MSA-English text from condition B/C.

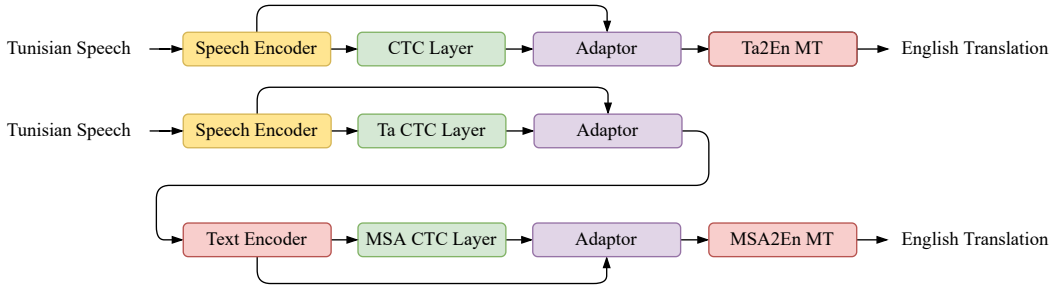

Figure 2: The top figure shows the SATE model (Xu et al., 2021), which implements a forward dialect transfer system from MSA to Tunisian through pre-training and fine-tuning techniques. The bottom part shows the Hybrid SATE model with a hierarchical text encoder, which can be used to reversely transfer from Tunisian to MSA.

3.3 End-to-end ST Model

The end-to-end ST approaches can mitigate issues of error propagation that often appears in low-resource scenarios. We developed an E2E ST system utilizing the SATE model (Xu et al., 2021) due to its effectiveness and simplicity for implementation, which is shown in Figure 2. In particular, we suggest two dialect transfer approaches for condition B/C, specifically the forward dialect transfer system from MSA to Tunisian and the reverse dialect transfer method from Tunisian to MSA.

3.3.1 Forward dialect transfer system

The forward dialect transfer system aims to transfer information from MSA to Tunisian by pre-training the ASR and MT models on the MSA dataset, respectively. These models are then fine-tuned using the Tunisian dataset to transfer from MSA to Tunisian. Note that the forward dialect transfer system is treated as a transfer of model parameters. In order to create an E2E ST system, we utilize the SATE model with pre-trained Tunisian ASR and MT models, followed by fine-tuning the SATE model with Tunisian ST dataset.

During training, the SATE model utilizes multi-task optimization, including the CTC loss of the source language $\mathcal{L}_{\text{CTC}}^{\text{Ta}}$, the cross-entropy loss for the target language $\mathcal{L}_{\text{CE}}^{\text{En}}$ and the knowledge distillation (KD) losses for both the source and target languages, i.e., $\mathcal{L}_{\text{KD}}^{\text{Ta}}$ and $\mathcal{L}_{\text{KD}}^{\text{En}}$. The overall loss function reads

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CTC}}^{\text{Ta}} + \lambda_2 \mathcal{L}_{\text{CE}}^{\text{En}} + \lambda_3 \mathcal{L}_{\text{KD}}^{\text{Ta}} + \lambda_4 \mathcal{L}_{\text{KD}}^{\text{En}}, \quad (1)$$

with four respective hyper weight parameters. The SATE model utilizes an adaptor to map speech features into the text feature space but suffers from inconsistent in-between sequence lengths. For this, we proposed a robust training method. Specifically, the Tunisian ASR model was first decoded by retaining both the repeated tokens and blank symbols of the CTC output. The resulting output was then combined with its corresponding English text to fine-tune the Ta2En MT model. The modified Ta2En MT model was well-suited to initialize the MT module of the SATE model.

3.3.2 Reverse dialect transfer system

It is a common issue that the Tunisian Arabic dialect is considered as being non-standardized at the linguistic level (Ben Abdallah et al., 2020). To address this, we proposed a reverse dialect transfer system that converts the Tunisian dialect to MSA, serving as a regularization of the dialect, which is illustrated in Figure 2. We modified the SATE model with a hierarchical text encoder (resulting in **Hybrid SATE**) to enable the reverse dialect transfer system. The proposed Hybrid SATE model primarily comprises a speech encoder, a Ta2MSA text encoder and an MSA2En MT module.

In order to initialize the model parameter for the Ta2MSA text encoder module in the Hybrid SATE model, we trained a Ta2MSA MT model. Based on the generated Ta-MSA* data in condition A and Ta**-MSA paired data in condition B/C from Section 3.2, we first pre-trained a Ta2MSA MT model with the Ta**-MSA data from condition B/C. Notably, the Ta2MSA MT model is equipped with a CTC layer on top of its encoder and is trained with an additional CTC loss for MSA. Then, we fine-tuned the model using the Ta-MSA* data from condition A. Finally, the encoder attached with a CTC layer of the Ta2MSA MT model was used to initialize the Ta2MSA text encoder.

The hybrid SATE model is optimized with an additional CTC loss for MSA, denoted as $\mathcal{L}_{\text{CTC}}^{\text{MSA}}$, resulting in the overall loss function

$$\begin{aligned}\mathcal{L} =& \lambda_1 \mathcal{L}_{\text{CTC}}^{\text{Ta}} + \lambda_2 \mathcal{L}_{\text{CE}}^{\text{En}} + \lambda_3 \mathcal{L}_{\text{KD}}^{\text{Ta}} + \lambda_4 \mathcal{L}_{\text{KD}}^{\text{En}} \\ &+ \lambda_5 \mathcal{L}_{\text{CTC}}^{\text{MSA}}. \end{aligned} \quad (2)$$

3.4 Model Ensemble Method

As training a single model can lead to implicit model bias, it is expected that a model ensemble decoding method can improve system robustness, especially in low-resource ST scenarios. We implemented synchronous decoding with multiple models and averaged the posterior probabilities predicted by each model at each time step. Consistent with single model decoding, the beam search decoding strategy was used with a beam size of 10. Subsequently, multiple models decoded the next tokens based on the same historical tokens. It should be noted that either E2E ST or MT models can be used for the model ensemble. Consequently, we can form ensembles of E2E ST and cascaded ST systems by using transcriptions from the ASR models as inputs for the MT models.

4 Experiments and results

4.1 Model Configurarions

ASR: For condition A, we employed the base model configurations, whereas the large model configurations were used for the experiments on condition B/C. Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) subword segmentation with the Tunisian text was trained and the dictionary size was 1000. The detailed model configurations are given in Appendix A.

MT: We considered two encoder-decoder architectures for MT: the normal transformer model (Vaswani et al., 2017) and the macaron-like transformer model (Lu et al., 2019). The latter uses several FFN-attention-FFN layers instead of the attention-FFN layer used in the former. Our MT model has three variants based on the number of layers in the encoder and decoder and the type of model architecture: MT base, MT large, and MT macaron. For detailed model and dictionary sizes, please refer to Table 13 in Appendix A.

E2E ST: Since both the SATE and hybrid SATE models are initialized by pre-trained ASR and MT modules, the model parameters can be inferred straightforwardly from the aforementioned ASR and MT model settings.

4.2 Results

4.2.1 Automatic Speech Recognition

Table 4 shows the ASR performance in terms of word error rate (WER) of MSA. Among the three different model structures, the VGG-Conformer model achieves the best performance. It is clear that the performance can be further improved by using additional private data in condition C.

Model	B		C	
	dev	test	dev	test
VGG-Conformer	14.3	13.2	12.5	12
VGG-Transformer	16.6	15.5	14.2	13.3
GateCNN-Conformer	15.1	14.2	14.3	13.4

Table 4: The **WER** of the MSA MGB2 corpus.

Model	A		B		C	
	dev	test1	dev	test1	dev	test1
VGG-Conformer	48.5	55.4	45.4	53.2	42	49.7
VGG-Transformer	49.2	57	49	56.8	44.7	52.1
GateCNN-Conformer	46.6	53.4	47.2	53.7	46.1	53.3
Ensemble	**44.5**	**51.7**	**43.4**	**50.9**	**40.8**	**48.7**

Table 5: The **original WER** on Tunisian. Due to the non-standard orthography and grammar in Tunisian, the value of original WER is relatively higher than the normalized WER in Table 11.

The pre-trained MSA ASR models are fine-tuned using Tunisian data for dialect transfer in condition B/C. As shown in Table 5, the VGG-Conformer model continues to perform best among different single models in condition B/C, while the GateCNN-Conformer model performs best in condition A. We further ensemble the three single models mentioned above and get the final ASR model results for each condition[3]. This demonstrates that model ensemble can significantly improve the ASR performance, especially in condition A. Comparing the ASR results in condition B/C with that in condition A, we find that pre-training on high-resource MSA data can improve the ASR performance in low-resource Tunisian.

4.2.2 Cascaded Speech Translation

We will demonstrate the usage of the BTFT data via an ablation study on condition A. For condition B/C, we compare the quality of different versions of Ta-En pseudo data. Besides, we introduce two methods for robust training, called **constrained fine-tune** and **error adaptation fine-tune**.

BTFT and Constrained Fine-Tune Our baseline MT model of condition A is trained using the original Ta-En MT data. From Table 6, we see

[3]For model ensemble of condition B, the VGG Transformer and GateCNN-Conformer models are from condition A, and the VGG-Conformer model is from condition B.

Data & Method	MT		Cascaded ST	
	dev	test1	dev	test1
Baseline	26.3	23.0	19.4	16.7
BTFT data	28.2	24.0	20.3	17.1
+ Constrained FT	**28.5**	**24.3**	**20.6**	**17.3**

Table 6: The BLEU score of MT and cascaded MT experiments in condition A.

Model	Pretrain Model	MT BLEU	
		dev	test1
En2Ta	-	12.4	10.0
En2Ta	En2MSA	16.6	12.5
MSA2Ta*	-	8.3	6.8
MSA*2Ta	MSA2Ta*	12.1	9.6

Table 7: The BLEU score of different pivot MT models using Ta-MSA*-En triple text data of condition A.

that combining the training data with BTFT data brings a considerable performance gain for both MT and cascaded ST. The MT model trained by the BTFT data are further fine-tuned by the original true paired Ta-En data. In order to prevent excessive over-fitting while fine-tuning, we proposed a constrained fine-tune method, as depicted in Figure 3. Specifically, the student model is constrained by the teacher model using KL divergence loss to avoid catastrophic forgetting and over-fitting. In case of using the constrained fine-tune method, the MT training objective function is given by

$$\mathcal{L} = \mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{CE}}. \quad (3)$$

Pseudo Ta-En paired data From Table 7, we see that the model initialized by a pre-trained model generates higher quality translations, i.e., higher quality pseudo-data. However, the performance comparison between the En2Ta model and the MSA*2Ta model may not be convincing since the input for the two models is different.

Comparing the performance of the Ta2En MT model is more appropriate to directly reveal the quality of the two versions of pseudo Ta-En data. In Table 8, it is clear that pre-training the MT model using Ta-En pseudo-data performs better than using MSA-En data. Moreover, the second version of Ta-En pseudo data outperforms the first when used for pre-training the Ta2En MT model. We believe that the MSA2Ta model is preferable for the En2Ta model due to the consistent use of MSA data during training and decoding. The En2Ta model employs English text from condition A for training, but uses English text from condition B/C to generate pseudo-Tunisian text. In comparison, the MSA2Ta model consistently uses MSA data from condition B/C for both training and decoding.

Model	MT		Cascaded ST	
	dev	test1	dev	test1
MSA2En-large	-	-	-	-
+ BTFT data FT	29.3	26.0	22.2	19.0
+ Constrained FT	30.1	26.2	22.5	19.2
Ta*2En-large	16.3	15.6	13.3	11.4
+ BTFT data FT	29.9	26.5	22.5	19.3
+ Constrained FT	30.4	26.6	22.8	19.5
Ta**2En-large	16.7	15.5	13.3	12.0
+ BTFT data FT	30.4	26.6	23.1	19.2
+ Constrained FT	**30.8**	**27.0**	**23.2**	**19.5**

Table 8: The BLEU score of the MT and the cascaded ST systems in condition C.

Model	MT		Cascaded ST	
	dev	test1	dev	test1
Condition A Best	28.5	24.3	20.6	17.3
+ Error Adapation FT	28.3	23.9	20.5	17.1
Condition C Best	30.8	27.0	23.2	19.5
+ Error Adapation FT	30.7	26.6	23.3	19.7

Table 9: The BLEU score of the MT and the cascaded ST systems in condition A/C when using error adaption fine-tune method.

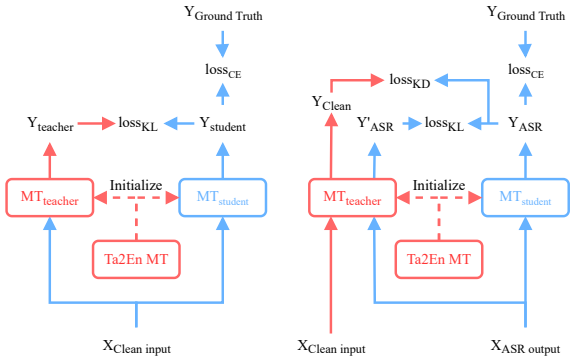

Figure 3: Left: Constrained Fine-tune, Right: Error Adaptation Fine-tune.

Error Adaptation Fine-tune As shown in Figure 3, the error adaptation fine-tune method (Zhang et al., 2022) slightly adjusts the MT model to mitigate potential ASR prediction errors. This technique fine-tunes the Ta2En MT model using a combination of the ASR output text and the text from the target language. It is based on the constrained fine-tune method by incorporating true text from

Model		SATE			Hybrid-SATE	Ensemble
Speech encoder		Conformer		Transformer	Conformer	
MT module		MT	MT-Macaron	MT	MT	
A	dev	20.2	20.1	19.5	-	21.2
	test1	17.2	17.3	16.6	-	18.2
B	dev	22.0	22.0	20.9	22.0	23.4
	test1	19.0	19.1	18.0	18.9	20.3
C	dev	23.8	23.7	23.4	23.1	24.9
	test1	20.7	20.2	20.0	20.2	22.0

Table 10: The BLEU scores of our E2E ST in condition A/B/C, where the speech encoder and MT module represent the sub-modules, and MT and MT-Macaron represent MT large and MT macaron models, respectively.

the source language as soft-labels to enhance the training with the KD loss $\mathcal{L}_{\text{KD}}$. The loss function for the error adaptation fine-tune method is given by

$$\mathcal{L} = 0.5\mathcal{L}_{\text{KD}} + 0.5\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CE}}. \quad (4)$$

From Table 9, we can observe that the error adaption fine-tune method enhances the performance of the cascaded ST system, albeit at a cost of MT performance decline. This reveals that this method is not effective in condition A but rather useful in condition B/C.

4.2.3 End-to-end Speech Translation

The SATE model can be instantiated in various structures by using different speech encoder and MT modules. Table 10 demonstrates that the conformer encoder outperforms the transformer encoder, showing an average improvement of 0.7 BLEU in condition A/B/C. For the different MT modules, the normal MT module is slightly better than the MT module in the macaroon form. Again, the results indicate model ensemble increases about 1.1 BLEU on the test1 set in condition A/B/C. The results of dialect transfer show an improvement for ST by 2.1 BLEU in condition B compared to condition A, and this is even greater in condition C, i.e., 3.8 BLEU. Additionally, the hybrid SATE model significantly improves the ST performance when used as a sub-model for model ensemble.

4.2.4 Model Ensemble

Table 11 presents the overall results of our ASR/MT/ST systems. The ASR results in terms of the normalized WER are derived from the model ensemble method in Table 5. It is worth noting that the ASR models are trained on original transcriptions but evaluated in a normalized form, which may cause a performance drop. The ensemble of three single MT models achieves an average improvement of 0.4 BLEU in text translation and cascaded ST systems of condition A/B/C, compared to the best single model of each data condition. The results of the E2E ST systems are derived from Table 10. We find that the E2E ST system falls

#	data condition	A	B	C
	ASR		WER↓	
	JHU-IWSLT2022	44.8	43.8	44.5
A1	**ASR Ensemble**	43.0	42.9	40.6
	MT		BLEU↑	
	CMU-IWSLT2022	22.8	23.6	-
M1	MT base	23.8	26.5	26.5
M2	MT large	23.9	26.3	26.6
M3	MT macaron	23.8	26.6	26.9
M4	**MT Ensemble**	24.3	26.9	27.4
	Cascaded ST		BLEU↑	
	CMU-IWSLT2022	17.5	17.9	-
C1	A1 + M1	17.7	19.3	19.6
C2	A1 + M2	17.8	19.5	20.0
C3	A1 + M3	17.6	19.5	19.9
C4	**A1 + M4**	18.4	19.9	20.2
	E2E ST		BLEU↑	
	CMU-IWSLT2022 (Mix)	18.7	18.9	-
E1	Ensemble of SATE	18.2	20.0	21.3
E2	**Ensemble of SATE + Hybrid SATE**	-	20.3	22.0
	Cascaded and E2E ST		BLEU↑	
	CMU-IWSLT2022 (Ensemble)	19.2	19.5	-
E3	Ensemble of C4 + E1	19.0	20.5	21.4
E4	**Ensemble of C4 + E2**	-	20.8	21.9

Table 11: The overall results of our ASR/MT/ST systems on **test1** set. The hypothesis and reference are normalized before computing **normalized WER** in order to be consistent with last year's ASR system. We substituted the MT base model of condition C with the MT base model of condition B. JHU-IWSLT2022 and CMU-IWLST2022 are taken from (Yang et al., 2022) and (Yan et al., 2022), respectively.

slightly behind the cascaded system in condition A but significantly surpasses it in condition B/C.

In the constrained condition, the primary system of our submission comprises an ensemble of cascaded and E2E ST models (see row **E3** of condition A). Additionally, for the unconstrained condition, we add the hybrid SATE model to the ensemble of cascaded and E2E ST models, which leads to a significant improvement of approximately 0.4 BLEU. Although the ensemble of cascaded and E2E ST system shows a 0.1 BLEU drop in condition C, it helps achieve the best performance in condition A/B. Therefore, the primary system of the submission for the unconstrained condition is in row **E4** of condition C. Moreover, we submit a contrastive system (i.e., row **E4** of condition B) to compare the performance without using private data.

5 Conclusion

This paper presents the methods and experimental results of the USTC team for the dialect ST (Tunisian Arabic to English) task in IWSLT 2023. The proposed forward and reverse dialect transfer methods, which were shown to be effective for augmenting text data and building hybrid SATE models. We utilized various model structures for implementing ASR, MT and ST tasks, and improved the robustness through model ensembling and error adaptation during training. The experiments showed a significant improvement in dialectal ST through the use of dialect transfer method. In unconstrained condition, our E2E ST system performs better than the cascaded ST system but is slightly less effective in constrained condition. Future studies might include the exploration of E2E ST models for unified modeling of multiple dialects (e.g., Tunisian, Egyptian) with MSA.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62101523), Hefei Municipal Natural Science Foundation (2022012) and USTC Research Funds of the Double First-Class Initiative (YD2100002008). We would like to thank Zhongyi Ye, Xinyuan Zhou and Ziqiang Zhang for valuable discussions and also thank Kevin Duh, Paul McNamee and Kenton Murray for organizing Tunisian Arabic to English track of the dialectal and low-resource speech translation shared task.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estève, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutail Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Najla Ben Abdallah, Saméh Kchaou, and Fethi Bougares. 2020. Text and speech-based Tunisian Arabic sub-dialects identification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6405–6411, Marseille, France. European Language Resources Association.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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 *Proc. Interspeech 2020*, pages 5036–5040.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*.

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*.

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. In *Proc. Interspeech 2019*, pages 2613–2617.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

Chen Xu, Bojie Hu, Yanyang Li, Yuhao Zhang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2021. Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2619–2630, Online. Association for Computational Linguistics.

Brian Yan, Patrick Fernandes, Siddharth Dalmia, Jiatong Shi, Yifan Peng, Dan Berrebbi, Xinyi Wang, Graham Neubig, and Shinji Watanabe. 2022. CMU's IWSLT 2022 dialect speech translation system. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 298–307, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur. 2022. JHU IWSLT 2022 dialect speech translation system description. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 319–326, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Weitai Zhang, Zhongyi Ye, Haitao Tang, Xiaoxi Li, Xinyuan Zhou, Jing Yang, Jianwei Cui, Pan Deng, Mohan Shi, Yifan Song, Dan Liu, Junhua Liu, and Lirong Dai. 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, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

A Appendix. Model configurations

The detailed model configurations for ASR systems are as following:

- **Condition A:** The model configurations are almost identical to the ESPnet (Inaguma et al., 2020) baseline. There are 12-layer encoder and 6-layer decoder. The attention module of both the encoder and decoder comprises 256 hidden units and 4 attention heads. The size of the FFN module is 1024 for the encoder but 2048 for the decoder. We use two VGG blocks as the feature extractor for both the VGG-Conformer and the VGG-Transformer models. For the GateCNN-Conformer model, the feature extractor has a 6-layer GateCNN.

- **Condition B/C:** The model difference between the condition A and the condition B/C lies in the model size. For condition B/C, the attention module has 512 hidden units and 8 attention heads, and the size of FFN is 4096.

Condition	Training Stage	lr	Max-tokens	Warmup	Dropout rate	Training steps
A	Stage1: BTFT Pretrain	5e-4	12000	4000	0.3	120000
	Stage2: Constrained Fine-tune	-	4096	-	0.3	40000
B/C	Stage1: MSA-En Pretrain	1e-3	40000×8	4000	0.1	200000
	Stage2: Ta**-En Pretrain	5e-4	40000×8	None	0.1	20000
	Stage3: BTFT Fine-tune	4e-5	6144	4000	0.3	120000
	Stage4: Constrained Fine-tune	-	2048	-	0.3	80000
	Stage5: Error Adaptation Fine-tune	1e-5	4096	None	0.3	10000

Table 12: Hyper parameters in different stages ("-" means reuse from the former stage and "×" the GPU numbers).

Condition	**A**	**B/C**
Encoder dim	256	512
Encoder FFN dim	1024	2048
Encoder attn heads	4	8
Decoder dim	256	512
Decoder FFN dim	1024	2048
Decoder attn heads	4	8
Tunisian BPE units	1000	1000
MSA BPE units	-	32000
English BPE units	4000	32000

Table 13: The model sizes and dictionary sizes for MT training, where "attn" represents attention module.

#	A	B	C
test2		ASR WER↓	
IWSLT2022	43.8	42.9	41.5
A1	**40.8**	**40.5**	**39.3**
test2		ST BLEU↑	
IWSLT2022	20.4	20.8	18.7
E3	**20.5**	-	-
E4	-	**22.8**	**23.6**
test3		ASR WER↓	
A1	43.2	42.3	40.5
test3		ST BLEU↑	
E3	18.1	-	-
E4	-	20.2	21.1

Table 14: The overall results of our ASR/ST systems on **test2** set (IWSLT 2022 evaluation set) and **test3** set (IWSLT 2023 evaluation set).

For MT models, the 6-layer encoder and 6-layer decoder are used for both MT base and MT macaron models, but 12-layer encoder and 6-layer decoder for MT large model. The details of the MT system are summarized in Table13.

B Appendix. Training and Inference

ASR: We used the fairseq tool (Ott et al., 2019) for training and inference. During training, we used a dropout rate of 0.3, set the label-smoothing rate to 0.1 and used a CTC loss weight of 0.3. The max tokens and max sentences per batch were 32000 and 120, respectively. We used the inverse square learning rate schedule for training, with a learning rate of 1e-3 and warmup steps of 8000 for condition A. For condition B/C, we pre-trained with MSA ASR data and used a learning rate of 1e-3 and warmup steps of 30000. We used a learning rate of 2e-4 and warmup steps of 8000 while fine-tuning with in-domain Tunisian ASR data. The models were optimized through the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.98$. During inference, we used an attention-based decoding strategy with a beam size of 10. We averaged the model parameters of 5 best model based on the WER on the dev set.

MT: The MT model training was also conducted using the fairseq toolkit. We conducted all training stages on the NVIDIA A40 GPU, varying the specific GPU number depending on the stage. Different training methods and hyper-parameters were used for optimal results depending on the condition, where we classified them into condition A and B/C. Specifically, we divided our training method into several stages, see Table 12. In Stage2 and Stage5 of condition B/C, the number of training steps is significantly lower than other stages. This was because the model had a tendency to overfit quickly during these stages; hence learning rate warmup method was not used during training. During inference, the beam size of decoding is 10. We used the official sacrebleu tool (Post, 2018) to calculate the normalized case-insensitive BLEU score. We averaged the model parameters of 5 best models based on the BLEU score on the dev set.

E2E ST: The hyper-parameters of the model training and inference are almost consistent with those used for ASR. The knowledge distillation weight (KD) for ASR is set to 0.2 but 0.3 for MT. The CTC loss weight for the speech encoder is set

to 0.2 while it is 1.2 for the Ta2MSA text encoder of hybrid SATE. Note that the CTC loss weight for the Ta2MSA text encoder is much larger because translating Tunisian to MSA with pseudo Ta-MSA MT data is challenging.

C Appendix. Official Evaluation Results

The official evaluation results of our submitted systems on both test2 and test3 sets (both being blind tests) are summarized in Table 14. Our submissions outperformed last year's best performance in all data conditions (constrained and unconstrained) for both ASR and ST evaluations (e.g, see the results of test2 set).