JHU IWSLT 2023 Dialect Speech Translation System Description

Amir Hussein[†] **Cihan Xiao**[†] **Neha Verma**[†] **Thomas Thebaud**[†]
Matthew Wiesner[‡] **Sanjeev Khudanpur**[†‡]
[†]Center for Language and Speech Processing, and
[‡] Human Language Technology Center of Excellence,
Johns Hopkins University
{ahussei6, cxiao7, nverma7, tthebau1, wiesner, khudanpur}@jhu.edu

Abstract

This paper presents JHU's submissions to the IWSLT 2023 dialectal and low-resource track of Tunisian Arabic to English speech translation. The Tunisian dialect lacks formal orthography and abundant training data, making it challenging to develop effective speech translation (ST) systems. To address these challenges, we explore the integration of large pre-trained machine translation (MT) models, such as mBART and NLLB-200 in both end-to-end (E2E) and cascaded speech translation (ST) systems. We also improve the performance of automatic speech recognition (ASR) through the use of pseudo-labeling data augmentation and channel matching on telephone data. Finally, we combine our E2E and cascaded ST systems with Minimum Bayes-Risk decoding. Our combined system achieves a BLEU score of 21.6 and 19.1 on test2 and test3, respectively.

1 Introduction

The performance of machine translation systems is closely tied to the amount of available training data. Regional dialects, which are less prevalent and primarily spoken languages, pose a challenge for these systems due to the scarcity of digital data, the absence of standard orthography, and prevalence of non-standard grammar. The IWSLT 2023 dialect and low-resource track focuses these challenges.

In this paper we present the JHU Tunisian Arabic to English speech translation systems submitted to the IWSLT 2023 dialectal and low-resource track (Agarwal et al., 2023). Arabic and its dialects form a *dialect continuum* anchored by Modern Standard Arabic (MSA) (Badawi et al., 2013). While MSA is the language of *formal* and *written* communication, most native Arabic speakers colloquially use local *dialects*, which often lack a standardized written form. In many North African Arabic dialects, including Tunisian, there is a significant code-switching with and borrowing from several *contact languages*: Berber and Romance languages like French, Spanish and Italian.

Recent successes in machine translation (MT) of text for low-resource languages or non-standard dialects have entailed the use of large pretrained models such as mBART (Liu et al., 2020a) and NLLB (NLLB Team et al., 2022). These models have demonstrated state-of-the-art performance via transfer learning from higher-resource languages, particularly through related languages. However, there is a lack of understanding regarding how to effectively integrate these models with speech recognition systems to develop speech translation systems. To fill this gap we investigate dialect transfer by integrating large pretrained models with speech recognition models in end-to-end (E2E) and cascaded speech translation (ST) systems. The key components of our system are:

- Dialectal transfer from large pre-trained models to improve translation in both E2E and Cascaded ST settings (§3.1,§3.2).
- Improved ASR of dialectal speech by reducing orthographic variation in training transcripts, and by channel matching (§3.1.1).
- System combination with Minimum Bayes-Risk decoding based on the COMET similarity metric (§3.3).

Our system outperforms the best previous approaches (Yang et al., 2022; Yan et al., 2022) for both ASR (WER) and ST (BLEU). We also found that integrating pre-trained MT models into end-to-end ST systems did not improve performance.

2 Dialect Speech Translation Task

The dialect speech translation task permitted submissions using models trained under two data conditions, (A) constrained and (B) unconstrained. For

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 283–290
July 13-14, 2023 ©2023 Association for Computational Linguistics

Condition	ASR	MT
(A) Basic	166 hours of manually transcribed Tunisian telephone speech	212K lines of manual English translation of the Tunisian transcripts
(B) Unconstrained	1200 hours of Modern Standard Arabic broadcast speech (MGB-2) (Ali et al., 2016). 250 hours of Levantine Arabic telephone conversations (LDC2006S29[1], LDC2006T07[2])	Any other English, Arabic dialects, or multilingual models beyond English and Arabic

Table 1: Data used for constrained and unconstrained conditions.

brevity, we will refer to these conditions as (A) and (B) respectively.

2.1 Data description

The data we used for the conditions (A) and (B) are listed in Table 1, and sizes of the training, development-testing and test partitions are listed in Table 2. The development and test sets for Tunisian data are provided by the organizers of IWLST 2023. The data is 3-way parallel: Tunisian Arabic transcripts and English translations are available for each Tunisian Arabic audio utterance. We use the development set for model comparison and hyperparameter tuning, and the test1 set for evaluating our ST systems. Finally, the task organizers provided blind evaluation (test2, test3) sets for final comparison of submissions.

	ASR (hours)	MT (lines)
train (condition A)	160	~202k
train (condition B)	1200+160+250	-
dev	3.0	3833
test1	3.3	4204
test2	3.6	4288
test3	3.5	4284

Table 2: Details for train, dev and test1 sets for constrained condition (A) and unconstrained condition (B).

3 Methods

In this section we describe our cascaded (§3.1), and end-to-end (E2E) (§3.2) speech translation systems as well as our strategy for combining both approaches (§3.3).

3.1 Cascaded ASR-MT

3.1.1 Automatic Speech Recognition

To train ASR models for E2E and cascaded systems, we use the ESPnet (Watanabe et al., 2018) toolkit. Our ASR architecture uses a Branchformer encoder (Peng et al., 2022), a Transformer decoder (Vaswani et al., 2017) and follows the hybrid CTC/attention (Watanabe et al., 2017) approach. Each Branchformer encoder block consists of two branches that work in parallel. One branch uses self-attention to capture long-range dependencies while the other branch uses a multi-layer perceptron with convolutional gating (Sakuma et al., 2021) to capture local dependencies. To mitigate orthographic variations (or inconsistencies) in the ASR transcripts, we augment the training data during the fine-tuning stage by reusing the audio training samples paired with their *ASR transcripts*, which tend to be orthographically more consistent. We refer to this approach as *pseudo-labeling*.

Condition (A). We train the ASR model described previously using the constrained Tunisian Arabic audio and transcripts.

Condition (B). The ASR Branchformer in this condition is pretrained on our MGB-2 standard Arabic data (Ali et al., 2016) and then fine-tuned on the provided Tunisian Arabic data. The MGB-2 MSA data differ from the Tunisian data in channel, and dialect. Since the Tunisian data are telephone conversations sampled at 8kHz, we downsample the MGB-2 speech from 16kHz to 8kHz, which we previously found was more effective than upsampling the telephone conversations to 16kHz (Yang et al., 2022). We also added additional telephone speech from the Levantine Arabic dialect (Maamouri et al., 2006). Note that Levantine Arabic is very different from Tunisian, and the hope here is to benefit from matched genre and channel conditions, not dialect.

We did not explicitly attempt to reduce the dialect mismatch. However, we mitigated some of the spurious orthographic variations in transcripts of dialectal speech by using pseudo-labels for training instead of of the manual transcripts, as noted above, in the final fine-tuning step.

3.1.2 Machine Translation

Condition (A). We train an MT model on Tunisian Arabic transcripts paired with their English translations. The MT architecture is similar to §3.1.1 model architecture, and uses a Branchformer encoder and Transformer decoder.

[1] https://catalog.ldc.upenn.edu/LDC2006S29

[2] https://catalog.ldc.upenn.edu/LDC2006T07

Condition (B). We experiment with two main pre-trained models: mBART and NLLB-200. In the first setting, we use the mBART25 model, which was shown to be slightly better for MSA versus the newer mBART50 model (Liu et al., 2020a; Tang et al., 2020). mBART25 also contains French, Turkish, Italian, and Spanish, all of which contribute loanwords to Tunisian (Zribi et al., 2014). Although these loanwords are transcribed in the Arabic script in our data, there is prior evidence that multilingual language models can benefit from cross-lingual transfer even between different scripts of the same language (Pires et al., 2019).

For NLLB-200, we use the distilled 1.3 billion parameter version of the model, due to space constraints. This model is a dense Transformer distilled from the original NLLB-200 model, which is a 54 billion parameter Mixture-of-Experts model that can translate into and out-of 200 different languages. We note that this model supports Tunisian Arabic, the aforementioned contact languages, MSA, as well as other closely related Maghrebi dialects (Moroccan, Egyptian, Maltese). The breadth of language coverage seen during the training of NLLB-200 makes this model an attractive choice for a dialect speech translation task.

We fine-tune these models on the provided $\sim$ 200K lines of Tunisian Arabic-English data. The source side is normalized as described in Section 4. We preprocess all data with the provided pre-trained `sentencepiece` vocabularies released with the models with no pre-tokenization. Results on MT systems are included in Table 8.

3.2 End-to-End Speech Translation

For the constrained condition we adopt the hierarchical multi-decoder architecture proposed by (Yan et al., 2022).

Condition (A). The system consists of a multi-task learning approach, which combines ASR and MT sub-nets into one differentiable E2E system where the hidden representation of the speech decoder is fed as input to the MT encoder. Additionally, the authors proposed using a hierarchical MT encoder with an auxiliary connectionist temporal classification (CTC) loss on top of the speech encoder. The MT decoder performs cross-attention over both the speech encoder and MT encoder representations. The ASR module is initialized with a Branchformer trained on the Tunisian data. In this part, we explore the effect of text normalization on the E2E-ST system and pre-trained MT initialization.

Condition (B). For the unconstrained condition, we propose a novel E2E-ST system that incorporates the combination of a pretrained ASR module and a pretrained MT module. Specifically, we combine the Branchformer ASR module described in Section 3.1, with mBART (Liu et al., 2020b), which was fine-tuned on Tunisian data. We modify the ESPnet ST recipe to incorporate the mBART model trained by the fairseq (Ott et al., 2019) framework. The architecture of the model is shown in Figure 1. In contrast to the modified Hierarchical Multi-Decoder architecture for Condition (A), to fully exploit the effect of MT pretraining, we removed the speech attention from the MT decoder that attends to the hierarchical encoder's hidden representations.

Specifically, the ASR encoder module in the proposed architecture takes in a sequence of audio features $x_1, x_2, \cdots, x_T$ and generates a sequence of hidden representations with length N, optimized with respect to the ASR CTC objective. The ASR decoder takes in the ASR encoder's hidden representations and autoregressively produces a sequence of logits with length L trained by the label-smoothing loss. The hierarchical speech encoder module is trained directly by the ST CTC loss for generating auxiliary frame-level labels in the target language to aid the ST decoding process. The primary innovation of the proposed system lies in the fully-connected layer that maps the ASR decoder's output hidden representations to some representations that resemble mBART's encoder's embedding layer's outputs, making the full system differentiable. The ST encoder subsequently encodes the input representations and feeds them into its decoder. The ST decoder, slightly different from the vanilla mBART decoder, optionally runs hybrid/joint CTC decoding at inference time, with the ST-CTC auxiliary labels and the autoregressively generated ST outputs with target length M, i.e. $y_1^{ST}, y_2^{ST}, \cdots, y_M^{ST}$.

3.3 System Combination

We perform a system combination across 5 of our systems: best constrained end-to-end system, best unconstrained end-to-end system, best cascaded system, and 2 additional cascaded systems (Fernandes et al., 2022). The two additional systems use the ASR produced by our end-to-end systems,

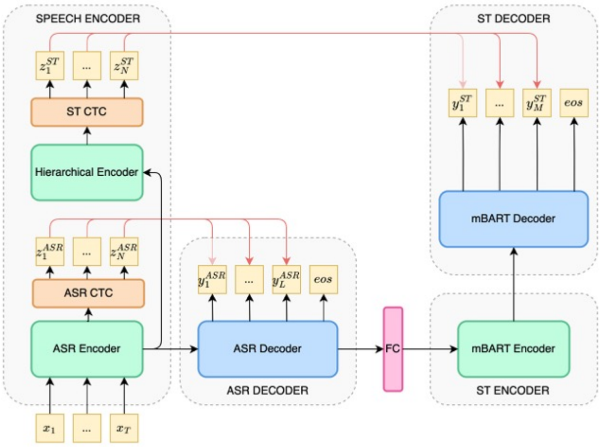

Figure 1: E2E model architecture with mBART MT module. The fully-connected (FC) layer applies a linear transformation to the ASR decoder's final hidden representation, which is then used to replace mBART's encoder's embedding layer's output.

and the same NLLB-200 MT component as in our best cascaded system. In Table 6, the 5 combined systems are referred to as A3, B1, B3, B4, and B5, in order.

3.3.1 Minimum Bayes Risk

We applied Minimum Bayes Risk decoding (Kumar and Byrne, 2004) to combine the hypotheses produced by five systems. For a given speech utterance x_i, and for a given system $s^j_{\theta_j}$ ($j \in \mathcal{S}$ and θ_j the set of parameters used by the j^{th} trained system), we can define the translation hypothesis as $y^j_i = f^j_{\theta_j}(x_i)$ and p^j_i be the probability that the hypothesis y^j_i would be outputted. We use this probability as a self-confidence score. Let $\mathcal{L}$ be similarity metric used to compare two hypothesis, outputting a scalar that rises if the two hypothesis are more similar. Then, for a given speech utterance x_i, and for a given set of systems $\mathcal{S}$, we define the best output as the one minimizing the distance with others while having the highest confidence:

$$y^{mbr}_i = \max_{y^j_i} \sum_{j \in \mathcal{F}} p^j_i \sum_{k \in \mathcal{F}} \mathcal{L}(y^j_i, y^k_i) \qquad (1)$$

3.3.2 Variations of MBR

Baseline MBR For our first combination, we compute the outputs according to the MBR using the BLEU score of sacrebleu (Post, 2018a) as the $\mathcal{L}$ similarity metric and the posterior probabilities p^j_i used are the log-likelihood ratios given by the end-to-end systems and the MT systems.

Unscored MBR For our second combination, we use the same technique but with a constant $p^j_i = 1$ for every system, as a simplified version of the Generalized MBR (Duh et al., 2011).

COMET-MBR For our third combination, we utilized the comet-mbr framework, which employs the COMET score between the source and hypothesis as the similarity metric ($\mathcal{L}$), using same equation (1), without the use of posterior probabilities (Fernandes et al., 2022). We used wmt20-comet-da for MBR scoring (Rei et al., 2020). Despite Tunisian Arabic not being a COMET-supported language, we observed an improvement compared to our single best system, suggesting that this approach may extend to dialects of languages covered by COMET.

4 Experiments

In this section, we describe our experiments on the ASR, MT, and ST tasks. In order to reduce the orthographic variation in the Tunisian speech transcription we performed additional text normalization similar to (Yang et al., 2022) which showed significant improvements on ASR, MT and ST tasks. The normalization is performed on both Tunisian and MSA transcripts and includes removing diacritics and single character words, and Alif/Ya/Ta-Marbuta normalization (see (Yang et al., 2022) for more details).

4.1 ASR

First we augment the raw audio segments by applying speed perturbation with three speed factors of 0.9, 1.0 and 1.1 (Ko et al., 2015). Then we transform the augmented audio to a sequence of 83-dimensional feature frames for the E2E model;

80-dimensional log-mel filterbank coefficients with 3 pitch features (Ghahremani et al., 2014). We normalize the features by the mean and the standard deviation calculated on the entire training set. In addition, we augment the features with specaugment approach (Park et al., 2019), with mask parameters $(mT, mF, T, F) = (5, 2, 27, 0.05)$ and bi-cubic time-warping. The E2E Branchformer-based ASR model was trained using Adam optimizer for 50 epochs with dropout-rate 0.001, warmup-steps of 25000 for condition (A) and 40000 for condition (B). The BPE vocabulary size is 500 for condition (A) and 2000 for condition (B). Table 3 summarizes the best set of parameters that were found for the Branchformer architecture. We note here that the Branchformer has 28.28 M parameters, which is approximately one-fourth the number of parameters in the Conformer (Yang et al., 2022), which is 116.15 M.

Att heads	CNN	Enc layers	Dec layers	d^k	FF
4	31	12	6	256	2048

Table 3: Values of condition (A) and (B) hyperparameters CNN: refers to CNN module kernel, Att: attention, Enc: encoder, Dec: decoder, and FF: fully connected layer

MGB2-tune: the pretrained model on MGB-2 is fine-tuned on Tunisian data from condition (A) by updating all model parameters with $1/10$ of the learning rate that was used during the training similar to (Hussein et al., 2021). In addition, we examine the effect of adding ASR outputs to the ground truth source during finetuning (**pseudo labeling**) and adding additional telephone data (**Tel**). The ASR results are summarized in Table 4 and compared to the state-of-the-art conformer results from (Yang et al., 2022). The MD refers to the hierarchical multi-decoder ST architecture adopted from (Yan et al., 2022), and MD-ASR refers to the ASR sub-module of the ST. It can be observed that the Branchformer provides slightly better results compared to the previous best conformer with similar size on both conditions (A) and (B). In addition, it can be also seen that pseudo labeling provides 2% relative improvement. We found that there is a high inconsistency between different transcribers since there is no standard orthography in Tunisian dialect. By incorporating the ASR predictions in this way, we aim to provide the model with more examples of the Tunisian dialect and help it better generalize to variations in the spoken language. To confirm this hypothesis we take a closer look at the most frequent top four substitutions shown in Table 5. The words are transliterated using Buckwalter transliteration (BW)[3] to make it readable for non-Arabic speakers. It can be seen that the ASR substitutions are present in both hypothesis and as correct reference which indicates that the assumption of reference inconsistency holds true. Finally, channel matching using more telephone data provides an additional 2.5% relative improvement.

		dev	test1	test2	test3
ASR-ID	Model	WER (↓)			
A1	Conformer (Yang et al., 2022)	40.8	44.8	43.8	-
A2	Branchformer	**40.1**	**44.5**	-	-
B1	MGB2-tune (Yang et al., 2022)	38.8	43.8	42.8	-
B2	MGB2-tune Branchformer	38.3	43.1	-	-
B3	+ Pseudo	37.5	42.6	-	-
B4	+ Tel	**36.5**	**41.7**	**40.6**	**41.6**
B5	E2E-MD-ASR	40.6	45.1	43.7	44.9
B6	E2E-mBART-ASR	37.7	43.2	41.5	42.6

Table 4: WER (%) of ASR models on dev, test1, test2 and test3 sets. A* and B* IDs are the ASR models developed under condition (A) and condition (B) respectively. B5 refers to the ASR submodule of the MD-ASR system under the constrained condition and B6 refers to the ASR sub-module of the E2E-mBART system both described in Section 3.2.

BW (REF / HYP)	Arabic	English Translation
69: Ayh / Ay	اي / ايه	yes
61: Ay / Ayh	ايه / اي	yes
18: Akhw / khw	اكهو / كهو	it's
17: khw / Akhw	كهو / اكهو	it's
8: gdwA / gdwh	غدوه / غدوا	tomorrow
7: gdwh / gdwA	غدوا / غدوه	tomorrow

Table 5: Top 6 substitutions with inconsistencies for ASR system transliterated using Buckwalter (BW). The number of times each error occurs is followed by the word in the reference and the corresponding hypothesis.

4.2 MT

We train the MT models as described in Section 3.1.2. For condition (A) the MT system parameters are shown in Table 7. In this condition, our MT system is finetuned on the training Tunisian data where the source data is mixed with ASR outputs, in order to be more robust to noisy source data. We use 5000 Byte-pair encoding (BPE) units shared between Tunisian Arabic and English. We train

[3] https://en.wikipedia.org/wiki/Buckwalter_transliteration

		Pretrained		dev	test1	test2	test3
ST-ID	Type	ASR	MT	BLEU (↑)	BLEU (↑)	BLEU (↑)	BLEU (↑)
A1	Cascade	A2	A3	18.9	15.6	-	-
A2	E2E-MD (Yan et al., 2022)	A2	-	20.6	17.1	-	-
A3	E2E-MD+norm	A2	-	**20.7**	**17.5**	19.1	17.6
B1	E2E-mBART	B4	B2	20.7	17.5	17.5	17.1
B2	Cascade-mBART	B4	B2	20.9	17.9	-	-
B3	Cascade-Base-NLLB200	B4	B3	**22.2**	**19.2**	**21.2**	**18.7**
B4	Cascade-B5-ASR-NLLB200	B5	B3	21.1	18.3	19.9	18.2
B5	Cascade-B6-ASR-NLLB200	B6	B3	22.2	18.8	20.7	18.3
B6	MBR with scores	-	-	21.7	18.8	18.7	17.1
B7	MBR no scores	-	-	22.7	19.6	20.6	18.8
B8	`comet-mbr`	-	-	**22.7**	**19.6**	**21.6**	**19.1**

Table 6: Results of cascaded, E2E, and combined systems measured by BLEU score on the dev, test1, test2 and test3. E2E-MD is the hierarchical multi-decoder described in (§3.2). Norm indicates the use of text normalization (§4) which is used with all systems except A2. The pretrained indicates the use of pretrained ASR and MT systems from Tables(8,4). A* and B* IDs are the models developed under condition (A) and condition (B) respectively

	layers	embed-dim	FF-embed	att-heads
Encoder	6	256	1024	4
Decoder	6	256	2048	4

Table 7: Values of constrained MT system parameters Enc: encoder, Dec: decoder, and FF: feed-forward

			dev	test1
MT-ID	Model Type	Model Size	BLEU (↑)	BLEU (↑)
A1	Transformer (Yang et al., 2022)		24.5	21.5
A2	Transformer Espnet	13.63 M	23.5	19.9
A3	Branchformer Espnet	16.81 M	25.0	21.4
B1	Transformer (Yang et al., 2022)		29.0	25.0
B2	mBART	610M	29.2	24.6
B3	NLLB-200	1.3B	**30.5**	**26.4**

Table 8: BLEU scores of various MT models using the gold reference transcripts. A* and B* IDs are the MT models developed under condition (A) and condition (B) respectively.

with the Adam optimizer; the maximum learning rate is 3e-03, attained after 20000 warm-up steps, and then decayed according to an inverse square root scheduler; we use dropout probability of 0.3; the model is trained for 200 epochs. For condition (B), for both NLLB-200 and mBART25, we finetune our model for up to 80000 updates and use loss to select our best model checkpoint. We use `sacrebleu` to compute the case-insensitive BLEU scores for all evaluation sets (Papineni et al., 2002; Post, 2018b) as shown in Table 8. The comparative analysis of our Espnet MT transformer with the best MT models reported in previous works based on Fairseq transformer (Yang et al., 2022) reveals a noticeable performance lag of up to -1.6 in absolute BLEU. However, incorporating the Branchformer module yields similar performance to the best Fairseq model. Finally finetuning NLLB-200 MT achieves the best results in the unconstrained category with 30.5 and 26.4 BLEU scores.

4.3 ST

Table 6 presents the results of our submitted cascaded and E2E ST systems. The pretrained column refers to the pretrained ASR and MT systems from Tables (4, 8). B1 denotes the end-to-end ST with B4 ASR and B2 mBART under the unconstrained condition, as described in Section 3.2. The E2E-MD is a hierarchical multi-decoder architecture described in Section 3.2, where the MT component is trained from scratch. The cascaded ST systems, Cascade-Base-NLLB200, Cascade-B5-ASR-NLLB200 and Cascade-B6-ASR-NLLB200, utilize the best MT model (NLLB200 B3) and ASR submodules including branchformer (B4), branchformer finetuned in E2E-MD setup (B5) and branchformer finetuned in with mBART setup (B6) respectively from Table 4.

It can be seen that the E2E-multidecoder architecture outperforms the cascaded system in the constrained condition, with a significant improvement of up to +1.7 in absolute BLEU. Text normalization provides additional boost of +0.4 in absolute BLEU. On the other hand for the unconstrained system, we observe that the cascaded system B2 outperforms the E2E B1 by up to 0.4 in absolute BLEU that utilizes identical submodules. The reason for this performance difference may be attributed to the inability of the input linear layer that was added to the MT encoder in the E2E setup (B1) to adjust the length of the ASR output to match the length of the mBART encoder's tokenization. This length discrepancy may lead to a loss of crucial information during the integration of the two modules, ultimately resulting in a degradation of overall performance. Further analysis is required to confirm this hypothesis and to identify potential solutions to address this issue. The highest performance of single ST system is obtained using Cascade-NLLB200-1.3B with BLEU of 21.2 and 18.7 on test2 and test3 respectively. Finally, we combine A3, B1, B3, B4 and B5 with `comet-mbr` which achieves the highest BLEU scores of 21.6 and 19.1 on test2 and test3 respectively.

5 Conclusion

In this paper, we have presented our submission for the IWSLT 2023 dialect speech translation task. We compared end-to-end to cascaded systems under constrained and unconstrained conditions. We found that an E2E-ST system outperformed the cascaded system under the constrained condition, while the cascaded models significantly outperformed the E2E-ST systems under the unconstrained condition. We provided a new E2E-ST baseline combining large pretrained MT with ASR under the unconstrained condition. Finally, we demonstrated that pseudo-labeling and channel matching provided significant improvements for the ASR and hence improved cascaded ST systems. In future work we plan to explore more effective ways of integrating the large pretrained MT models into E2E ST systems.

Acknowledgements

This work was partially supported by NSF CCRI Grant No 2120435. In addition, we thank the Qatar Computing Research Institute, Doha, for providing some of the computational resources that made this work possible.

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

El Said Badawi, Michael Carter, and Adrian Gully. 2013. *Modern written Arabic: A comprehensive grammar*. Routledge.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2011. Generalized minimum bayes risk system combination. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1356–1360.

Patrick Fernandes, António Farinhas, Ricardo Rei, José De Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2494–2498. IEEE.

Amir Hussein, Shammur Chowdhury, and Ahmed Ali. 2021. Kari: Kanari/qcri's end-to-end systems for the interspeech 2021 indian languages code-switching challenge. *arXiv preprint arXiv:2106.05885*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation.

Mohamed Maamouri et al. 2006. Levantine arabic qt training data set 5, speech ldc2006s29. Web Download.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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 NAACL-HLT 2019: Demonstrations*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

Matt Post. 2018b. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Jin Sakuma, Tatsuya Komatsu, and Robin Scheibler. 2021. Mlp-based architecture with variable length input for automatic speech recognition.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Yalta, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. *ArXiv*, abs/1804.00015.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11:1240–1253.

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.

Inès Zribi, Rahma Boujelbane, Abir Masmoudi, Mariem Ellouze, Lamia Belguith, and Nizar Habash. 2014. A conventional orthography for Tunisian Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2355–2361, Reykjavik, Iceland. European Language Resources Association (ELRA).