Edinburgh's End-to-End Multilingual Speech Translation System for IWSLT 2021

Biao Zhang[1] **Rico Sennrich**[2,1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computational Linguistics, University of Zurich
B.Zhang@ed.ac.uk, sennrich@cl.uzh.ch

Abstract

This paper describes Edinburgh's submissions to the IWSLT2021 multilingual speech translation (ST) task. We aim at improving multilingual translation and zero-shot performance in the constrained setting (without using any extra training data) through methods that encourage transfer learning and larger capacity modeling with advanced neural components. We build our end-to-end multilingual ST model based on Transformer, integrating techniques including adaptive speech feature selection, language-specific modeling, multi-task learning, deep and big Transformer, sparsified linear attention and root mean square layer normalization. We adopt data augmentation using machine translation models for ST which converts the zero-shot problem into a zero-resource one. Experimental results show that these methods deliver substantial improvements, surpassing the official baseline by > 15 average BLEU and outperforming our cascading system by > 2 average BLEU. Our final submission achieves competitive performance (runner up).[1]

1 Introduction

Although end-to-end (E2E) speech translation (ST) has achieved great success in recent years, outperforming its cascading counterpart and delivering state-of-the-art performance on several benchmarks (Ansari et al., 2020; Zhang et al., 2020a; Zhao et al., 2020), it still suffers from the relatively low amounts of dedicated speech-to-translation parallel training data (Salesky et al., 2021). In text-based machine translation (MT), one solution to lack of training data is to jointly perform multilingual translation with the benefit of transferring knowledge across similar languages and to low-resource directions, and even enabling zero-shot translation, i.e. direct translation between language pairs unseen in training (Firat et al., 2016; Johnson et al., 2017). However, whether and how to obtain similar success in very low-resource (and practical) scenario for multilingual ST with E2E models remains an open question.

To address this question, we participated in the IWSLT2021 multilingual speech translation task, which focuses on low-resource ST language pairs in a multilingual setup. Apart from *supervised* evaluation, the task also offers *zero-shot* condition with a particular emphasis where only automatic speech recognition (ASR) training data is provided for some languages (without any direct ST parallel data). The task is organized in two settings: *constrained* setting and *unconstrained* setting. The former restricts participants to use the given multilingual TEDx data (Salesky et al., 2021) alone for experiment; while the latter allows for additional ASR/ST/MT/others training data. In this paper, we address the constrained one.

Our E2E multilingual ST model takes Transformer (Vaswani et al., 2017) as the backbone, and follows the adaptive feature selection (AFS) framework (Zhang et al., 2020a,b) as shown in Figure 1. AFS is capable of filtering out uninformative speech features contributing little to ASR, effectively reducing speech redundancy and improving ST performance (Zhang et al., 2020a). We adapt AFS to multilingual ST, and further incorporate several techniques that encourage transfer learning and larger capacity modeling, ranging from language-specific modeling, multi-task learning, deep and big Transformer, sparsified linear attention (ReLA) (Zhang et al., 2021b) to root mean square layer normalization (RMSNorm) (Zhang and Sennrich, 2019b). Inspired by Zhang et al. (2020c), we convert the zero-shot translation problem into a zero-resource one via data augmentation with multilingual MT models.

[1]Source code and pretrained models are available at https://github.com/bzhangGo/zero.

Proceedings of the 18th International Conference on Spoken Language Translation, pages 160–168
Bangkok, Thailand (Online), August 5–6, 2021.

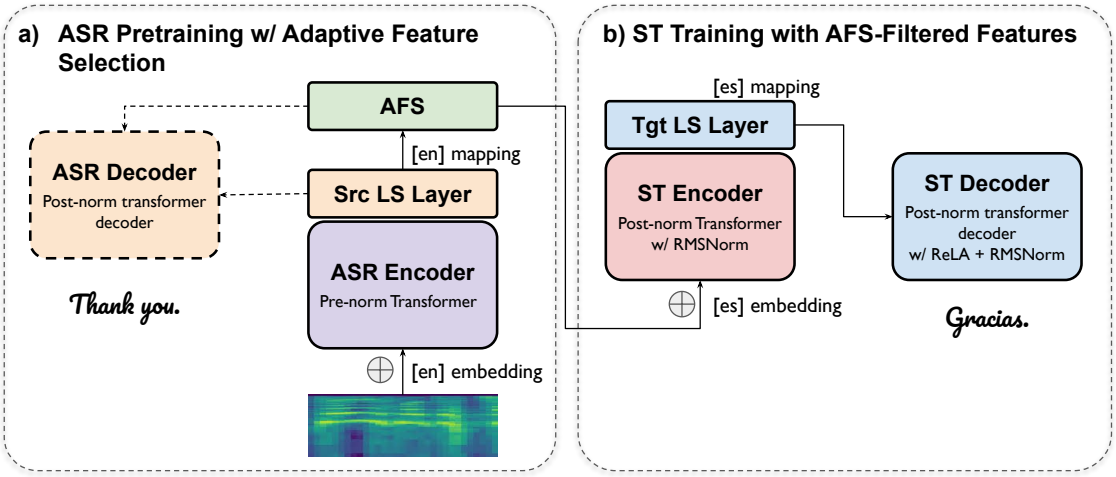

Figure 1: Overview of our multilingual ST model for an English-Spanish example. We first pretrain the ASR encoder paired with adaptive feature selection (AFS) to induce informative speech features (a), which are then carried over to the ST encoder-decoder model for translation (b). We adopt language embedding and language-specific (LS) linear mapping before and after ASR/ST encoder, respectively, to strengthen source/target (Src/Tgt) language modeling. The ASR decoder is discarded and the other ASR modules are frozen after the pretraining. Solid arrows illustrate the E2E translation procedure.

We integrate all these methods into one model for our submission. Our results reveal that:

- These methods are complementary in improving translation performance, where data augmentation and larger-capacity modeling contribute a lot.
- Low-resource E2E ST benefits greatly from multilingual modeling; our E2E multilingual ST performs very well in this task, outperforming its cascading counterpart by 2 average BLEU.

2 Methods

In this section, we elaborate crucial ingredients in our E2E multilingual ST, which individually have already been proven successful for ST or (multilingual) MT. We put them together to improve multilingual ST as shown in Figure 1. Note all encoder/decoder modules are based on Transformer (Vaswani et al., 2017).

2.1 Adaptive Feature Selection

Speech is lengthy and noisy compared to its text transcription. Also, information in an audio often distributes unevenly. All these increase the difficulty of extracting informative speech features. To solve this issue, researchers resort to methods compressing and grouping speech features (Salesky et al., 2019; Gaido et al., 2021). Particularly, Zhang et al. (2020a) propose adaptive feature selection (AFS) to sparsify speech encodings by pruning out those uninformative ones contributing little to ASR based on L_0DROP (Zhang et al., 2020b). Using AFS, Zhang et al. (2020a) observe significant performance improvements (>1 BLEU) with the removal of $\sim$84% speech features on bilingual ST.

Our model follows the AFS framework, which includes three steps: 1) pretraining the ASR encoder-decoder model; then 2) finetuning the ASR model with AFS; and 3) training ST model with the ASR encoder and the AFS module frozen.

2.2 Deep Transformer Modeling

Neural models often benefit from increased modeling capacity, and one way to achieve this is to deepen the models (He et al., 2015; Zhang et al., 2020d). However, simply increasing model depth for Transformer results in optimization failure, caused by gradient vanishing (Zhang et al., 2019a). To enable deep Transformer, Zhang et al. (2019a) propose depth-scaled initialization (DS-Init) that only requires changing parameter initialization without any architectural modification. DS-Init successfully helps to train up to 30-layer Transformer, substantially improving bilingual and also massively multilingual translation (Zhang et al., 2019a, 2020c). We adopt this strategy for all deep Transformer experiments.

Apart from DS-Init, researchers also find that changing the post-norm structure to its pre-norm alternative improves Transformer's robustness to deep modeling, albeit slightly reducing quality (Wang et al., 2019; Zhang et al., 2019a). We

keep using post-norm Transformer for most modules but apply the pre-norm structure to the ASR encoder to stabilize the encoding of speeches from different languages.

2.3 Language-Specific Modeling

Analogous to multi-task learning, multilingual translation benefits from inter-task transfer learning but suffers from task interference. How to balance between shared modeling and language-specific (LS) modeling so as to maximize the transfer effect and avoid the interference remains challenging. A recent study suggests that scheduling language-specific modeling to top and/or bottom encoder/decoder sub-layers benefits translation the most (Zhang et al., 2021a), resonating with the findings of Zhang et al. (2020c). In particular, Zhang et al. (2020c) propose language-aware linear transformation, a language-specific linear mapping inserted in-between the encoder and the decoder which greatly improves massively multilingual translation.

We adopt such language-specific linear mapping and apply it to both ASR and ST encoders. We ground such modeling in the ASR and ST encoder to the source and target language, respectively. Following multilingual translation (Johnson et al., 2017; Gangi et al., 2019; Inaguma et al., 2019), we adopt language embedding (such as "*[en]*, *[es]*") but add it to the inputs rather than appending an extra token.

2.4 Sparsified Linear Attention

Attention, as the key component in Transformer, takes the main responsibility to capture token-wise dependencies. However, not all tokens are semantically correlated, inspiring follow-up studies on sparsified attention that could explicitly zero-out some attention probabilities (Peters et al., 2019; Zhang et al., 2021b). Recently, Zhang et al. (2021b) propose rectified linear attention (ReLA) which directly induces sparse structures by enforcing ReLU activation on the attention logits. ReLA has achieved comparable performance on several MT tasks with the advantage of high computational efficiency against the sparsified softmax models (Peters et al., 2019).

Results on MT show that ReLA delivers better performance when applied to Transformer decoder (Zhang et al., 2021b). We follow this practice and apply it to the ST decoder. Our study also demonstrates that ReLA generalizes well to ST.

2.5 Root Mean Square Layer Normalization

Layer normalization (LayerNorm) stabilizes network activations and improves model performance (Ba et al., 2016), but raises non-negligible computational overheads reducing net efficiency, particularly to recurrent models (Zhang and Sennrich, 2019a). To overcome such overhead, Zhang and Sennrich (2019b) propose root mean square layer normalization (RMSNorm) which relies on root mean square statistic alone to regularize activations and is a drop-in replacement to LayerNorm. RMSNorm yields comparable performance to LayerNorm in a series of experiments (Zhang and Sennrich, 2019b) and show great scalability in large-scale pretraining (Narang et al., 2021).

We apply RMSNorm to the ST encoder and decoder, which benefits the training of deep and big Transformers.

2.6 Data Augmentation

Data augmentation (DA) is an effective strategy for low-resource tasks by increasing the training corpus with pseudo-labelled samples (Sennrich et al., 2016a; Zhang and Zong, 2016). Methods for generating such samples vary greatly, and we adopt the one following knowledge distillation (Kim and Rush, 2016). Note, prior to our study, knowledge distillation has already been successfully applied to ST tasks (Liu et al., 2019; Gaido et al., 2020). We regard the multilingual MT as the teacher since text-based translation is much easier than and almost upper-bounds the speech-based counterpart (Zhang et al., 2020a), and transfer its knowledge into our multilingual ST (student).

Concretely, we first train a multilingual MT model and then use it to translate each source transcript into all possible ST directions, including the zero-shot ones, based on beam search algorithm. We directly concatenate the generated pseudo speech-translation pairs with the original training corpus for multilingual ST training. This will convert the zero-shot translation problem into a zero-resource one for ST, which has been demonstrated effective in massively multilingual MT (Zhang et al., 2020c).

2.7 Multi-Task Learning

Multi-task learning aims at improving task performance by jointly modeling different tasks within one framework. Particularly, when tasks are of high correlation, they tend to benefit each other and de-

Speech	Target Languages				
	En	Es	Fr	Pt	It
Es	36K/102K	102K/-	3.6K/102K	21K/102K	5.6K/102K
Fr	30K/116K	21K/116K	116K/-	13K/116K	-/116K
Pt	31K/90K	-/90K	-/90K	90K/-	-/90K
It	-/50K	-/50K	-/50K	-/50K	50K/-

Table 1: Statistics for ST training data used for the IWSLT2021 multilingual ST task. "-": denotes no data available. "*a*/*b*": "*a*" denotes genuine data while "*b*" is for augmented data.

liver positive knowledge transfer. With datasets of different tasks combined, this also partially alleviates data scarcity.

We adopt multi-task learning by augmenting translation tasks with transcription tasks. We incorporate the ASR tasks for multilingual ST, and auto-encoding tasks (transcript-to-transcript in the same language) for multilingual MT.

3 Experimental Settings

In this section, we explain the used datasets, model architectures, optimization details and evaluation metrics in our experiments. All implementations are based on the *zero*[2] toolkit (Zhang et al., 2018).

Data We participate in the constrained setting, where only the provided data, i.e. Multilingual TEDx (Salesky et al., 2021), is permitted. Multilingual TEDx collects audios from TEDx talks in 8 source languages (Spanish/Es, French/Fr, Portuguese/Pt, Italian/It, Russian/Ru, Greek/El, Arabic/Ar, German/De) paired with their manual transcriptions, covering translations into 5 target languages (English/En, Es, Fr, Pt, It). It contains supervised training data for 13 ST directions, three of which (Pt-Es, It-En, It-Es) are masked-out for zero-shot evaluation. ASR training data is given for all 8 source languages. Overall, Multilingual TEDx is a small-scale dataset, whose ST training data size ranges from 5K utterances (It-Es) to at most 39K utterances (Es-En). Thus, studying and improving transfer across different languages is of great significance. The IWSLT2021 task requires participants to model translations from 4 source languages (Es, Fr, Pt, It), where the final evaluation only targets translations into En and Es. The statistics of ST (genuine and augmented) training data are shown in Table 1.

Regarding audio preprocessing, we use the given audio segmentation (train/dev/test) for experiments. We extract 40-dimensional log-Mel filterbanks with a step size of 10ms and window size of 25ms as the acoustic features, followed by feature expansion via second-order derivatives and mean-variance normalization. The final acoustic input is 360-dimensional, a concatenation of the features corresponding to three consecutive and non-overlapping frames. We tokenize and truecase all text data using Moses scripts (Koehn et al., 2007). We adopt subword processing (Sennrich et al., 2016b) with 8K merging operations (Sennrich and Zhang, 2019) on these texts to handle rare words. Note we use different subword models (but with the same vocabulary size) for ST, ASR and MT.

Architecture The architecture for ASR and ST is illustrated in Figure 1, while our MT model follows Zhang et al. (2020c). We apply AFS to ASR encoder outputs (after language-specific mapping) along both temporal and feature dimensions. By default, we adopt Transformer-base setting (Vaswani et al., 2017): we use 6 encoder/decoder layers and 8 attention heads with a model dimension of 512/2048. For deep Transformer, we equally increase the encoder and decoder depth, and adopt DS-Init for training. We also use Transformer-big for ST, where the number of attention heads and model dimension are doubled, increased to 16 and 1024/4096, respectively.

Optimization We train MT models with the maximum likelihood objective ($\mathcal{L}_{\text{MLE}}$). Apart from $\mathcal{L}_{\text{MLE}}$, we also incorporate the CTC loss (Graves et al., 2006) for ASR pretraining with a weight value of 0.3 following Zhang et al. (2020a). During AFS finetuning, the CTC loss is discarded and replaced with the L_0DROP sparsification loss (Zhang et al., 2020b) weighted by 0.5. We employ label smoothing of value 0.1 for $\mathcal{L}_{\text{MLE}}$.

We adopt Adam (β_1=0.9, β_2=0.98) for parameter tuning with a warmup step of 4K. We train all models (ASR, ST and MT) for 100K steps, and finetune AFS for 10K steps. We group instances of around 25K target subwords into one mini-batch. We apply dropout to attention weights and residual connections with a rate of 0.1 and 0.2, respectively. Dropout rate on residual connections is increased to 0.3 for ST big models to avoid overfitting, and to 0.5 for MT models inspired by low-resource MT (Sennrich and Zhang, 2019). Except dropout, we use *no* other regularization techniques. We use beam search for decoding, and set the beam size and length penalty to 4 and 0.6, separately. The

[2] https://github.com/bzhangGo/zero

Model	Es-En	Es-Pt	Es-Fr	Fr-En	Fr-Es	Fr-Pt	Pt-En	Pt-Es	It-En	It-Es	Avg
Bilingual Models⋆	25.5	39.3	2.0	28.3	30.5	19.0	27.9	29.9	18.9	1.0	22.23
Multilingual Models⋆	24.6	37.3	18.1	28.2	32.1	30.6	28.8	38.4	20.9	25.1	28.41
Our Multilingual MT											
+ 6 layers	28.7	42.1	29.3	33.6	38.3	36.7	33.2	42.9	20.3	32.7	33.78
+ 12 layers	31.8	44.7	31.7	36.4	40.9	39.9	35.6	44.0	23.0	34.9	36.29
+ 24 layers	32.8	44.9	32.4	37.3	41.8	40.7	36.8	43.2	23.2	35.3	**36.84**
Ablation Study											
+ 6 layers w/o LS layer	28.6	41.8	29.0	33.7	38.2	36.3	33.2	42.5	20.7	32.6	33.66
+ 6 layer + RoBT	28.1	40.3	28.6	34.1	38.3	33.6	33.6	42.7	21.1	32.9	33.33

Table 2: SacreBLEU↑ for MT on Multilingual TEDx testsets. ⋆: results reported by Salesky et al. (2021). Note the results for Pt-Es, It-En and It-Es translation in our model are based on zero-shot evaluation. In spite of this unfairness, our model still substantially outperforms the supervised baseline (Salesky et al., 2021) by a large margin, +8.43 BLEU. *RoBT*: random online back-translation (Zhang et al., 2020c). Best average BLEU is highlighted in **bold**. Columns in red denote zero-shot evaluation.

Model	Es	Fr	Pt	It	Ru	El	Ar	De	Avg
Hybrid LF-MMI⋆	16.2	19.4	20.2	16.4	28.4	25.0	80.8	42.3	**31.09**
Transformer⋆	46.4	45.6	54.8	48.0	74.7	109.5	104.4	111.1	74.31
Our Multilingual ASR									
+ 6 layers	17.6	19.5	23.1	20.8	39.8	33.0	104.3	57.8	**39.49**
Ablation Study									
+ 6 layers w/o LS layer	18.0	19.5	23.2	21.6	40.8	35.2	97.8	62.6	39.84

Table 3: WER↓ for ASR on Multilingual TEDx testsets. ⋆: results reported by Salesky et al. (2021). Best results are highlighted in **bold**.

model used for evaluation is averaged over the last 5 checkpoints.

Note, while the training data size varies across languages, we follow the original data distribution and adopt *no* specific sampling strategies for all multilingual experiments.

Evaluation We evaluate translation quality using tokenized case-sensitive (Sacre)BLEU (Papineni et al., 2002; Post, 2018), and report WER for ASR performance without punctuation on lowercased text. In ST experiments, we observe some repeated translations decreasing BLEU. We automatically post-process translations by removing repeated chunks of up to 10 words.

4 Results

4.1 Multilingual MT

Table 2 shows the results for text-based translation. Our best model, achieved with 24 layers, largely surpasses the official baseline (Salesky et al., 2021) by $>$ 8 average BLEU. With 6 layers, our model still largely surpasses this baseline by 5.37 average BLEU, suggesting the superiority of our model.

Increasing model depth greatly benefits multilingual MT (+2.51 average BLEU, 6 layers → 12 layers), even though the dataset is small. Note the benefit from increased depth diminishes as the depth goes larger (+0.55 average BLEU, 12 layers → 24 layers). We find that language-specific modeling slightly improves translation performance (+0.12 average BLEU). Such improvement seems uninteresting particularly compared to the significant gains on massively multilingual MT (Zhang et al., 2020c), but we ascribe this to the high language similarity in Multilingual TEDx and the relative small number of languages. We also confirm the effectiveness of random online back-translation (RoBT), which improves zero-shot translation via pseudo sentence pair augmentation (Zhang et al., 2020c). Table 2 shows that RoBT indeed benefits zero-shot translation, but sacrifices overall quality (-0.45 average BLEU).

Overall, our results reveal very positive transfer between these languages, and also great zero-shot translation performance. This is an encouraging finding for multilingual ST. We use our 24-layer model for data augmentation distillation in the following ST experiments.

Model	Es-En	Es-Pt	Es-Fr	Fr-En	Fr-Es	Fr-Pt	Pt-En	Pt-Es	It-En	It-Es	Avg
Multilingual Models*	12.3	17.4	6.1	12.0	13.6	13.2	12.0	13.7	10.7	13.1	12.41
Cascades with Multilingual MT*	21.5	26.5	23.4	25.3	26.9	23.3	22.3	26.3	21.9	28.4	24.58
Our Multilingual MT, w/ AFS, LS layer, DA, ReLA (decoder self-attention) and RMSNorm											
+ 6 layers	24.9	34.8	26.6	30.0	33.8	33.2	27.4	33.9	20.7	30.8	29.61
+ 12 layers	24.6	35.6	26.7	29.9	33.7	33.5	28.5	34.4	21.1	30.6	29.86
+ 6 layers + big model	26.1	36.2	27.5	31.0	34.9	34.3	28.7	35.1	21.6	31.5	**30.69**
Ablation Study											
+ 6 layers w/o AFS	25.2	35.1	26.4	29.9	33.2	32.7	28.4	33.7	20.3	29.6	29.45
+ 6 layers w/o AFS & DA	20.8	30.9	18.5	24.7	27.6	27.0	23.8	27.2	13.8	20.0	23.43
+ 6 layers w/o ReLA & RMSNorm	24.2	34.8	26.4	29.5	34.1	33.4	27.5	33.7	20.7	30.3	29.46
+ 6 layers + ReLA on cross-att.	24.8	35.3	27.1	30.2	34.3	33.8	27.6	34.1	20.5	30.5	**29.82**
Our Cascade Model w/ Multilingual ASR + 24-layer Multilingual MT											
	24.8	33.7	25.3	29.2	32.7	32.2	26.9	31.7	18.5	27.1	28.21
Final Submission: Ensemble of 4 base model, 1 12-layer model and 1 big model w/ length penalty of 0.9											
	26.6	36.6	27.9	31.8	35.6	35.4	29.7	35.8	22.0	32.0	**31.34**

Table 4: SacreBLEU↑ for ST on Multilingual TEDx testsets. *: results reported by Salesky et al. (2021). Note the results for Pt-Es, It-En and It-Es translation in our model are based on zero-shot evaluation. Our model substantially outperforms the official baseline (Salesky et al., 2021) by $>$ 10 average BLEU. *DA*: data augmentation. Best average BLEU is highlighted in **bold**.

4.2 Multilingual ASR

Table 3 shows the ASR performance. Following previous studies (Salesky et al., 2021; Zhang et al., 2020a), we experiment with the Transformer base setting. Our multilingual ASR model yields an average WER of 39.49, substantially outperforming the official baseline (Salesky et al., 2021) by 34.82 and narrowing the performance gap against the hybrid model to $\sim$ 8 WER. Note lower WER indicates better quality. We ascribe this large quality gain to the dedicated multilingual ASR model architecture, the better optimization, and particularly the incorporation of the CTC objective.

Removing the language-specific layer slightly hurts recognition performance (+0.35 average WER). It largely benefits ASR for Ar (-6.5 WER), but hurts that for De (+4.8 WER), showing the difficulty of multilingual modeling: it's hard to balance between different tasks (translation directions). We adopt the model with language-specific projection for AFS and ST.

Notice that we still include Ru, El, Ar and De for the ASR training, although they are not a part of the evaluation campaign. We regard this inclusion as some sort of model regularization: the extra training data could reduce overfitting and might enable potential cross-lingual transfer.

4.3 Multilingual ST

Table 4 summarizes the ST results. Our base model using 6 layers delivers an average BLEU of 29.61, largely outperforming the official baseline (Salesky et al., 2021) by $\sim$ 17 BLEU and also beating their cascading baseline. In a fair comparison where knowledge data augmentation is not used, our model still obtain an average BLEU of 23.43.

Increasing the ST model depth slightly improves quality (+0.25 average BLEU), while enlarging ST model yields a larger improvement, reaching 1.08. Although it's widely known that large neural model often suffers from overfitting in low-resource tasks, our results suggest that such model still gains quality with proper regularization (AFS, larger dropout, etc).

Our ablation study demonstrates the effectiveness of AFS, ReLA and RMSNorm, although the corresponding quality gains are marginal. In particular, we observe that applying ReLA to both self-attention and cross-attention in the ST decoder helps (Zhang et al., 2021b). AFS improves training efficiency, allowing larger batch size thus fewer gradient accumulation steps (Zhang et al., 2020a). Besides, data augmentation benefits multilingual ST very much, resulting in $\sim$ 6 average BLEU improvement, and the gain on zero-shot directions is even higher, + 7.54 BLEU. Thus, we mainly ascribe our success on zero-shot translation to the inclusion of pseudo parallel corpora – data matter! – which converts the zero-shot problem into a zero-resource one.

Our E2E model also largely outperforms the cascading system (+ 2.48 average BLEU). Notice that our cascading system is sub-optimal, since we

Model	Es-En	Es-Fr	Es-It	Es-Pt	Fr-En	Fr-Es	Fr-Pt	Pt-En	Pt-Es	It-En	It-Es	Avg
Ensemble of 6 E2E models: 4 base model, 1 12-layer model and 1 big model w/ length penalty of 0.9												
	36.2	30.3	32.9	44.5	26.4	29.5	30.1	27.0	34.5	23.0	31.1	**31.41**
Cascading model: base ASR model + 24-layer MT model												
	33.3	26.8	28.6	39.9	23.7	26.9	26.8	23.6	30.0	19.7	26.7	27.82
Single E2E Model: multilingual ST model + 6 layers, big Transformer												
	35.0	29.9	31.9	44.1	25.5	28.8	29.0	26.2	33.3	22.4	30.1	30.56

Table 5: SacreBLEU↑ for our submissions to the IWSLT2021 multilingual ST task.

didn't bias our MT model towards ASR outputs, and the mismatch between gold transcripts and ASR outputs often hurts cascading performance. Recent advances on avoiding such error propagation might deliver better cascading results (Cheng et al., 2018; Zhang et al., 2019b; Cheng et al., 2019; Sperber et al., 2019).

Our final submission is an ensemble of 6 E2E multilingual ST models, which reaches an average BLEU of 31.34. Apart from the ensemble, we also increase the decoding length penalty from 0.6 to 0.9, which performs slightly better.

5 Submission Results

The IWSLT2021 task prepares a held-out test set for the final evaluation. We submitted three systems: one cascading system, one E2E single model (w/ big ST Transformer) and one ensemble model. Results are shown in Table 5: our E2E multilingual ST model outperforms its cascading counterpart, and the ensemble model reaches the best performance. Our submission achieves runner-up results among all participants.

6 Conclusion and Future Work

We describe Edinburgh's end-to-end multilingual speech translation system for the IWSLT2021 multilingual speech translation task. We observe substantial performance improvement using larger-capacity modeling (deep or big modeling) and data augmentation. In spite of the scarcity of the training data, we show that E2E models benefit greatly from multilingual modeling and deliver promising results on zero-shot translation directions (even without data augmentation). Our E2E multilingual ST greatly surpasses its cascading counterpart.

Regarding future study, we argue that exploring the multilingual transfer behavior should be very practical and promising to ST. This work mainly studies transfer across similar languages. How the current model generalizes to distant languages is still an open question. Besides, a general trend for deep learning is to increase the model capacity via deep and/or big modeling. However, deep models for ST seem to be ineffective. Identifying the reason for this and proposing simple solutions would be of high interest.

Acknowledgements

We thank the reviewers for their insightful comments. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements 825460 (ELITR). Rico Sennrich acknowledges support of the Swiss National Science Foundation (MUTAMUR; no. 176727).

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv*, abs/1607.06450.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the data barrier: Towards robust speech translation via adversarial stability training. *CoRR*, abs/1909.11430.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. CTC-based compression for direct speech translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 690–696, Online. Association for Computational Linguistics.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-end speech-translation with knowledge distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 80–88, Online. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 585–592. IEEE.

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*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 570–577. IEEE.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. Interspeech 2019*, pages 1128–1132.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. 2021. Do transformer modifications transfer across implementations and applications?

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.

Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy. Association for Computational Linguistics.

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.

Elizabeth Salesky, Matthias Sperber, and Alan W Black. 2019. Exploring phoneme-level speech representations for end-to-end speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1841, Florence, Italy. Association for Computational Linguistics.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. The multilingual tedx corpus for speech recognition and translation.

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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Matthias Sperber, Graham Neubig, Ngoc-Quan Pham, and Alex Waibel. 2019. Self-attentional models for lattice inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1185–1197, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021a. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang and Rico Sennrich. 2019a. A lightweight recurrent network for sequence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1538–1548, Florence, Italy. Association for Computational Linguistics.

Biao Zhang and Rico Sennrich. 2019b. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020a. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544, Online. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2019a. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China. Association for Computational Linguistics.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2020b. On sparsifying encoder outputs in sequence-to-sequence models.

Biao Zhang, Ivan Titov, and Rico Sennrich. 2021b. Sparse attention with linear units.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020c. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2020d. Neural machine translation with deep attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):154–163.

Biao Zhang, Deyi Xiong, jinsong su, Qian Lin, and Huiji Zhang. 2018. Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4273–4283. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019b. Lattice transformer for speech translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6475–6484, Florence, Italy. Association for Computational Linguistics.

Chengqi Zhao, Mingxuan Wang, and Lei Li. 2020. Neurst: Neural speech translation toolkit.