ZJU's IWSLT 2021 Speech Translation System

Linlin Zhang
Zhejiang University
11921133@zju.edu.cn

Abstract

In this paper, we describe Zhejiang University's submission to the IWSLT2021 Multilingual Speech Translation Task. This task focuses on speech translation (ST) research across many non-English source languages. Participants can decide whether to work on constrained systems or unconstrained systems which can use external data. We create both cascaded and end-to-end speech translation constrained systems, using the provided data only. In the cascaded approach, we combine Conformer-based automatic speech recognition (ASR) with the Transformer-based neural machine translation (NMT). Our end-to-end direct speech translation systems use ASR pretrained encoder and multi-task decoders. The submitted systems are ensembled by different cascaded models.

1 Introduction

In this paper, we introduce our submission to the IWSLT2021 Multilingual Speech Translation Task. This task focuses on speech translation (ST) research across many non-English source languages. Multilingual models enable transfer from related tasks, which is particularly important for low-resource languages; however, parallel data between two otherwise high-resource languages can often be rare, making multilingual translation and zero-shot translation important for many resource settings. The task provides data for two conditions (Salesky et al., 2021): supervised, and zero-shot, including speech and transcripts for four languages (Spanish, French, Portuguese, Italian) and translations in a subset of five languages (English, Spanish, French, Portuguese, Italian). At evaluation time, using the provided speech in the four source languages, participants submit the generated translations in both English and Spanish.

In the cascaded approach, we use a Conformer (Gulati et al., 2020) model for ASR for every language. For the MT component, we use a unified Transformer model for all language pairs. As previous works (Gangi et al., 2019; Bahar et al., 2020), we use both the clean and noisy speech transcripts, back translation data, and the mask noisy trick.

For the end-to-end direct speech translation, we also created a Transformer-based model. To obtain the best possible translation quality, we apply data augmentation on audio files, make a multitask decoding for incorporating the ASR task (Weiss et al., 2017).

We tried various experimental parameter settings and different architectures, and finally submitted an ensembled cascaded system.

2 Cascaded Speech Translation

As the task provides speech and transcripts for four languages (Spanish, French, Portuguese, Italian) and translations in a subset of five languages (English, Spanish, French, Portuguese, Italian). Zero-shot language pairs have ASR data released for training but not translations. Cascades of separately trained automatic speech recognition and machine translation (MT) models can leverage all of these data sources.

2.1 Automatic Speech Recognition

We only focus on sequence-to-sequence ASR models. We firstly used a Transformer-based (Vaswani et al., 2017) model on FAIRSEQ[1]. Our transformer-based models presented as Synnaeve et al. (2019) consist of 2 1-D convolutional subsampler layers and 12 transformer encoder layers, 6 transformer decoder layers. The input mel-filterbank features are 80 dimensions, and the audio files' sample frequency is 16K. As Transformer models

[1] This tool can be found via https://github.com/pytorch/fairseq

Proceedings of the 18th International Conference on Spoken Language Translation, pages 144–148
Bangkok, Thailand (Online), August 5–6, 2021.

WER	layers	es	fr	pt	it
Transformer	12	15.68	17.23	21.69	20.66
	16	15.91	17.90	19.65	19.74
Conformer	12	15.1	16.7	18.8	18.9

Table 1: The results of the Transformer and Conformer ASR models with different encoder layers.

option	range
tempo	$(0.85, 1.25)$
speed	$(0.95, 1.05)$

Table 2: Sox parameters value ranges used in processing of audio data.

are good at capturing content-based global interactions, while CNNs exploit local features effectively. Then, we used the convolution-augmented transformer ASR model, Conformer (Gulati et al., 2020). The architecture settings are as the Conformer-base model using ESPnet[2] which also uses a joint CTC/attention decoding (Hori et al., 2017). The Conformer model consists of 2-D convolutional subsampler layers and 12 encoder layers, 6 decoder layers. The input features combine 80-dimension mel-filterbank features and 3 pitch features.

We remove all text sequences longer than 200 tokens and all speech utterances longer than 3000 frames. The two models both use a variant of SpecAugment(Park et al., 2019) for data augmentation. The Conformer model also used the speed perturbation technique. The results of the Conformer automatic speech recognition models are shown on the Table 1.

2.2 Multilingual Machine Translation

We created text-to-text machine translation baselines using FAIRSEQ (Ott et al., 2019a). We followed the recommended Transformer hyperparameters as the IWSLT'17 multilingual task. This model uses a shared BPE vocabulary of 16k learned jointly across all languages. We appended language ID tags to the beginning of each sentence for both the encoder and decoder.

For the provided translation data, some language pairs are zero shot. For example, language pair Italian-to-Spanish has no training data, but Spanish-to-Italian is provided. So we use the Spanish-to-Italian corpus in reverse and supplement it as the Italian-to-Spanish training corpus. The corpus of French to Spanish is also used in reverse, add to the training set of Spanish to French. This reverse use also adds language pairs, such as English-to-Spanish. At the same time, back translation (BT) is also used to generate a pseudo-corpus.

There is a gap between the transcription generated by the ASR model and the ground-truth transcription. In practice, the ASR-generated transcripts can be seen as noisy data by Gangi et al. (2019). We add the ASR-generated transcripts noisy data to train the MT model, to increase the system's robustness (Sperber et al., 2017).

At the same time, we also adopted the mask trick used in BERT (Devlin et al., 2019). We randomly mask some words in the source language sentence and use the last layer of encoder output to predict the masked words. The probability p of the masked tokens is 0.1.

We have not applied an individual bilingual translation model for each language pair while using a unified translation model for all language pairs. Our experiments show that multilingual text translation is more conducive to solving the zero-sample problem.

3 End-to-End Direct Speech Translation

We used FAIRSEQ to train end-to-end Transformer-based models for ST, using 80-dimensional mel-filterbank features with global Cepstral Mean and Variance Normalization (CMVN), SpecAugment (Park et al., 2019), and 1-D convolutions downsampler with the pretrained Transformer-based ASR model. We remove all text sequences longer than 200 tokens and all speech utterances longer than 6000 frames.

In order to make full use of the speech translation data of all language pairs, we adopt a joint vocabulary of 10K for all language pairs. In the beginning, we used the ASR model trained with all 4 languages ASR corpus to pre-train the ST, but in the end, the ASR model trained with just 1 language was used to pre-train the ST and the latter result was better. Same as the multilingual machine translation model, we prepend the source language ID tag to the frame sequence after the down-sampling of 1-D CNN layers. At the same time, we also prepend the target language ID tag to

[2] This tool can be accessed via https://github.com/espnet/espnet

	target				
source	en	es	fr	pt	it
es	39k(69h)	107k(189h)	7k(11h)	24k(42h)	6k(11h)
fr	33k(50h)	24k(38h)	119k(189h)	16k(25h)	-
pt	34k(59h)	zero-shot	-	93k(164h)	-
it	zero-shot	zero-shot	-	-	53k(107h)

Table 3: The number of sentences and the segment of audios for the Multilingual TEDx dataset. Same source and target languages mean the ASR data.

	ES-EN	FR-EN	FR-ES	PT-EN	PT-ES	IT-EN	IT-ES	
end-to-end	19.20	21.76	22.46	20.45	18.21	4.45	5.47	
+multi-task	19.61	22.69	23.45	21.20	20.79	4.31	5.83	
cascaded+(BT data)	24.01	28.52	33.67	28.07	36.52	15.21	27.04	MT
	20.29	24.51	26.83	22.42	26.71	14.61	22.13	ST
cascaded+(ASR noisy data)	25.11	30.16	34.14	29.13	36.69	15.42	26.62	MT
	20.56	24.60	26.81	22.03	26.46	14.58	22.07	ST
ensemble+beam12	21.28	26.21	28.98	23.43	27.99	15.71	23.19	ST

Table 4: The speech translation results of the test sets in BLEU score of different end-to-end and cascaded models.

the target text token sequence.

We augmented the data by processing the audio files with two Sox's effects as Potapczyk and Przybysz (2020): tempo, speed. We sampled the parameters with uniform distribution within ranges presented in Table 2: For each audio file, we repeated the process 2 times. The effect of this operation is basically similar to speed perturbation. Because ESPnet already uses speed perturbation by default, we only apply Sox's effects on the FAIRSEQ models.

As in many previous works, we also introduced a second decoder with ASR task, making it a multi-task setup similar to Weiss et al. (2017). The ASR and ST tasks use a joint dictionary of size 10k as the baseline. The training loss can be calculated as follows:

$$\text{Loss} = \text{Loss}_{ST} + \alpha\,\text{Loss}_{ASR} \tag{1}$$

We tried setting the value of α to 0.7 and 0.5, and the result was better when it was set to 0.5. Thus, the ASR and ST decoder are trained jointly, and convolutional layers and encoder are shared. The experiments also proved that this kind of multi-task learning is useful.

All the models consist of 12 encoder layers and 6 decoder layers, including the multi-task model.

For the one encoder-one decoder baseline, we just pretrain the encoder. For the multi-task model, we use the pretrained ASR model to initialize the shared encoder and ASR decoder. We also tried to pretrain only the shared encoder of the multi-task model. Our experimental results show that pretraining the ASR decoder will not improve the final effect of speech translation, but it can reduce the loss of the ASR decoder and the convergence time of training. We also tried to increase the number of encoder layers from 12 to 16, and the translation performance almost did not improve, but the number of convergence epochs decreased.

4 Experiments

In this section, we report the results for cascaded and end-to-end direct speech translation models on various data and settings.

For the ASR task, we tried 2 different platforms, the results as Table 1. For the cascaded speech translation models, the ASR part is implied on the ESPnet (Watanabe et al., 2018), while the MT component is implied on the FAIRSEQ (Ott et al., 2019b). For the end-to-end direct speech translation models, including the pretrained ASR models, all models are built on the FAIRSEQ.

For the cascaded speech translation models, all MT models have used the mask tokens trick, the main difference is just the different adding data. For the end-to-end direct speech translation models, all the models including the pretrained ASR models are trained including the Sox's effects data. All the parameter settings are almost unchanged. The ASR model trained with just 1 language (Spanish) was used to pretrain the ST. We tried only using Spanish or French ASR to pretrain the ST model, compared with all 4 multilingual ASR. Using multilingual ASR initialization led to a decrease of nearly 2.9 BLEU on ES-EN testset with only ES ASR. Pretraining with one language ASR is better than with all four languages, which surprised us a bit. We originally felt that the performance of the richer corpus model should be better. Perhaps understanding this problem will help improve the effectiveness of the multilingual end-to-end model.

Our multilingual translation model and end-to-end multilingual speech translation model both adopt a unified model for all language pairs, and do not apply special processing to individual language pairs.

4.1 Settings

For the Transformer-based ASR models are trained using the Adam optimizer, dropout probability of 0.1, and label smoothing. The learning rate schedule is inverse sqrt, with a learning rate 0.001, warmup from 10000. The same architecture is used to pretrain our direct speech translation models. The Conformer-based ASR model is also trained using the Adam optimizer and label smoothing, while warmup from 25000. For all ASR models, we apply byte-pair-encoding (BPE) (Sennrich et al., 2016) with 4k merge operations for every language.

For all the end-to-end direct ST models, the training settings are the same as the Transformer-based ASR models. While the multilingual end-to-end ST models apply BPE with 10k merge operations. All the models are trained of the 320000 batch size.

4.2 Results

As shown in Table 4, our cascade models represent better scores than our end-to-end models, particularly for low-resource language pairs. End-to-end models are closing the performance gap for high-resource settings. The early models on the experimental phase set the beam search size as 5 for saving time, while the final submitted ensemble model has a beam search size of 12. Finally, we submitted an ensembled cascaded system, which ensembles all multilingual MT models. The submitted model's BLEU scores are 34.5 on ES-EN, 25.2 on FR-EN, 27.4 on FR-ES, 25.7 on PT-EN, 31.6 on PT-ES, 20.8 on IT-EN, 27.3 on IT-ES.

References

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-before-end and end-to-end: Neural speech translation by apptek and RWTH aachen university. In *Proceedings of the 17th International Conference on Spoken Language Translation, IWSLT 2020, Online, July 9 - 10, 2020*, pages 44–54. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019. Robust neural machine translation for clean and noisy speech transcripts. *CoRR*, abs/1910.10238.

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 *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 5036–5040. ISCA.

Takaaki Hori, Shinji Watanabe, and John R. Hershey. 2017. Joint ctc/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 518–529. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019a. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019b. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. 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. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

Tomasz Potapczyk and Pawel Przybysz. 2020. Srpol's system for the IWSLT 2020 end-to-end speech translation task. In *Proceedings of the 17th International Conference on Spoken Language Translation, IWSLT 2020, Online, July 9 - 10, 2020*, pages 89–94. 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. *CoRR*, abs/2102.01757.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *International Workshop on Spoken Language Translation (IWSLT)*.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Edouard Grave, Tatiana Likhomanenko, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end ASR: from supervised to semi-supervised learning with modern architectures. *CoRR*, abs/1911.08460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2207–2211. ISCA.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA.