Multilingual Speech Translation KIT @ IWSLT2021

Ngoc-Quan Pham, Dan He, Tuan-Nam Nguyen,
Thanh-Le Ha, Sebastian Stüker, Alexander Waibel
Karlsruhe Institute of Technology
ngoc.pham@kit.edu

Abstract

This paper contains the description for the submission of Karlsruhe Institute of Technology (KIT) for the multilingual TEDx translation task in the IWSLT 2021 evaluation campaign. Our main approach is to develop both cascade and end-to-end systems and eventually combine them together to achieve the best possible results for this extremely low-resource setting. The report also confirms certain consistent architectural improvement added to the Transformer architecture, for all tasks: translation, transcription and speech translation.

1 Introduction

The neural sequence-to-sequence models have revolutionalised both automatic speech recognition (ASR) and machine translation in many different aspects, from performance (Luong et al., 2015; Pham et al., 2019a) to various forms such as multimodal (Barrault et al., 2018) and multilingual (Kannan et al., 2019; Ha et al., 2016; Johnson et al., 2016). After multilingual text translation has been established, the recent focus is naturally shifted to multilingual speech translation especially with a series of public speech corpora with multiple translation being released (Iranzo-Sánchez et al., 2020; Wang et al., 2020; Salesky et al., 2021).

Recent evaluation campaigns in speech translation have seen a fierce competition between traditional cascade systems and end-to-end counterparts (Jan et al., 2018, 2019; Ansari et al., 2020). The competition without a doubt would continue in multilingual speech translation especially in a low-resource condition. However, the competition between two modeling schemes suggests that each of them possesses its own strengths and advantages. Notably the cascade models can easily benefit from the separated optimized architectures of each subtask and enjoy the larger available datasets, while the end-to-end models can theoretically avoid *error propagation*.

This manuscript describes the translation system for the multilingual TEDx task with the aim of combining the strong points of both approaches. We showed that optimizing the cascade models is necessary to bootstrap a powerful end-to-end model, while in the end combining their powers based on ensembling gives promising results.

2 Dataset overview

The Multilingual TEDx corpus (Salesky et al., 2021) provided us with the 5 languages Spanish (es), French (fr), Italian (it), Portuguese (pt) and English (en). While speech audio is available for the first 4 languages, text translation is available for all 20 language pairs, and the speech translation parallel data is largely more scarce than the other two. The data statistics is shown in Table 1 and 2.

Source → Target	**en**	**es**	**fr**	**it**	**pt**
es	36K	102K	3.6K	5.6K	21K
fr	30K	20K	116K	-	-
it	-	-	-	50K	-
pt	-	30K	-	-	90K

Table 1: Data statistics for speech recognition/translation in the number of utterances.

Source → Target	**en**	**es**	**fr**	**it**	**pt**
en	-	36.2K	30.5K	-	30.8K
es	36.2K	-	24.4K	5.6K	21.1K
fr	30.1K	24.4K	-	-	13.2K
it	-	5.6K	-	-	-
pt	30.8K	21.1K	13.2K	-	

Table 2: Data statistics for machine translation in the number of sentence pairs.

It is noticeable that the training data is severely lacking for speech translation when the number of sentences is only a fraction of the ASR or MT resources. As a result, our initial plan was to generate

Proceedings of the 18th International Conference on Spoken Language Translation, pages 154–159
Bangkok, Thailand (Online), August 5–6, 2021.

synthetic translations from the available transcripts, that can effectively increase the data size for training end-to-end SLT models.

3 General enhancement for Transformer Models

In this section, we describe the overall model descriptions that were applied in all three tasks.

Transformers (Vaswani et al., 2017) are constructed with blocks of transformation functions including self-attention and feed-forward neural networks.

Self-attention transforms a sequence of states using themselves as queries, keys and values, building up hierarchical representational powers since the output states are the weighted-sum of the input states that can be flexibly learned during training. Relative attention (Shaw et al., 2018) further improves the interaction between states by assigning learnable weights for each relative position. (Pham et al., 2020) incorporated this mechanism into speech models by extending the partially learnable relative positions in (Dai et al., 2019) to attend to all positions in the sequence bidirectionally.

Furthermore, the Transformer models are strengthened by using dual feed-forward (FFN) layers per block instead of one (Lu et al., 2019). As such, one feed-forward network block precedes the initial self-attention in either encoder and decoder. The outputs of both FFN layers are scaled by 0.5. Besides, it is possible to help training deep Transformer better by using RELU-inspired activation functions that do not suffer from dead neurons. GELU (Hendrycks and Gimpel, 2016) and SiLU (Elfwing et al., 2018) are combined with gated linear units (Dauphin et al., 2017), as used in our activation functions.

In most of our experiments and in the eventual submission, all of the above enhancements were incorporated. Ablation studies are unfortunately not fully possible because of the time constraint, but will be provided to depict the improvement of each addition.

4 Speech Recognition

Our speech recognition models are built based on both the LSTM and the Speech Deep Transformer (Pham et al., 2019a) enhanced with bidirectional relative attention (Pham et al., 2020). While LSTM models have been intensively experimented for the best results (Nguyen et al., 2019a; Park et al., 2019), Transformers have been recently adopted to this task with strong results (Pham et al., 2019a, 2020).

For the four languages in the Multilingual TEDx, we trained both multilingual Transformers and LSTM models on the combination of the datasets, using the factorization scheme. The LSTM has 6 encoder layers and 2 decoder layers with 1024 hidden units in each layer. The sole attention layer between encoder and decoder is an 8-head dot-product attention. On the other hand, we experimented the Transformers with the "Large" models having 16 encoder layers and 6 decoder layers with 1024 units in the hidden layers.

The models are trained with Adam and an inverse square-root learning rate schedules with 4096 warm-up steps following the same setting as (Vaswani et al., 2017) for up-to 120K steps or early-stopping on the development set. In order to facilitate training, layers are randomly dropped with the highest rate of 0.5 and linearly reducing from top to bottom (Pham et al., 2019a). Due to the relatively small size of the dataset, regularization is added with dropout probability 0.35 in all layers, and spec augmentation with dropped frequency range is $F = 16$ and the maximum dropped time $T = 64$ which is relatively aggressive.

Language	LSTM	bTF	eTF	Ensemble
es	16.9	16.4	15.25	14.37
fr	16.5	16.8	15.39	14.44
pt	18.3	19.5	17.1	16.79
it	19.5	16.4	17.24	15.47

Table 3: Comparison on Multilingual TEDx dataset (WER↓). Our baseline models include the baseline (b) and enhanced (e) Transformers (TF) and the LSTM.

Table 3 shows the experimental result of speech recognition, in which we can see that the Transformer with only Relative attention is as good as the LSTM, while using all enhancements allowed us to improve the result further. It is notable that those results are obtained using our own word error rate measuring method that does not remove punctuations, which are retained in ASR to be compatible with the subsequent MT models.

Removing the punctuations and using the evaluation scripts in the same repo with (Salesky et al., 2021) gave us 11.0, 13.88, 13.38 and 14.14 error rates for Spanish, Italian, French and Portuguese respectively, which are significantly lower than the

Hybrid LF-MMI provided.

5 Machine Translation

Our multilingual machine translation is built based on the universal multilingual framework (Ha et al., 2016; Johnson et al., 2016; Pham et al., 2019b), in which the vocabulary is shared between languages using a BPE size of 16000 merging units.

Thanks to the relatively small data size, the translation task was used to measure the incremental improvement of various features, including the relative attention and the Macaron feed-forward layers. Therefore, experiments were carried out using the base-setting of Transformer as the starting point. Dropout was increased to 0.35 together with word dropout (Gal and Ghahramani, 2016) at both encoder and decoder to help the models counter overfitting. The output language is controlled by the language embedding vectors added directly to the word embedding at every timestep (Ha et al., 2017; Pham et al., 2019b). The language pairs are randomly sampled based on the training size of each pair (no temperature was used). Training is done using the adaptive learning rate for Adam, with maximum learning rate at 0.7 achieved after 4096 warming-up steps, and is often early-stopped after 60000 training steps, each is approximately 48000 words.

Regularization is further improved via data diversification (Nguyen et al., 2019b). Carrying a similar idea of back-translation (Sennrich et al., 2016) that generates synthetic labels for untranslated monolingual data, the main idea of data diversification is to popularize the available training data with synthetic translation of both source sentences and target sentences.

According to the algorithm presented in (Nguyen et al., 2019b), the training process is divided into rounds in which the training data is incrementally added with synthetic data coming from the refining models themselves. Starting from the original training data in round 0, we use the best settings in round n to translate the source and target sentences in the training to the counterpart language and add the synthetic translation pairs to the current training data, proceeding to round $n + 1$. Each synthetic pair consists of one original sentence and one synthetic sentence. The idea is the combination of backtranslation, model distillation (Kim and Rush, 2016) and data augmentation (Wang et al., 2018) without any additional data.

Interestingly, thanks to the multilingual property, it is also possible to translate one sentence to a range of languages after each round, leading to different options and a massive amount of sentences to be added. However, it was empirically found out that the method did not scale after 1 round, and massively translating to all languages did not improve the training data. Therefore, after round 0, the best configuration which is an ensemble is used to generate synthetic parallel data for round 1 by just translating each sentence to the same language in the original dataset.

The translation result is seen in Table 4. We showed the progressive results as a result of adding each empirical feature, and measured the change in average over 14 language pairs. Even though the training data also contains language pairs that are not included for the SLT task, we found that adding those "reverse" language pairs is beneficial for the others.

In terms of improvement, it can be seen that even though in this extreme low-resource scenario, using more complicated architecture obtained better translations. A combination of relative attention, macaron FFN and 16 layers of depth allowed us to improve the baseline by 0.95 BLEU points, in which the relative attention seems to be the most useful. Ensembling multiple models is, as expected but costly to improve the results further.

Data diversification was very effective after the first round, by improving the average score by nearly 1 BLEU point. Italian-related language pairs enjoyed up to 2 BLEU points, due to the lowest amount of original sentences. This result somewhat went against the initial expectation, because by not changing the sampling method, the data ratio for those languages was even lower than in round 0.

We obtained the best configuration for text translation with ensembles on round 1. Proceeding to round 2 unfortunately did not produce any further improvement, which might be reasoned by the dominance of synthetic sentences in terms of quantity.

6 End-to-end Speech Translation

Naturally, end-to-end speech translation is developed at the last stage to benefit from the previous stages. The ASR models serve as providing the SLT with the pretrained encoder, while we used the MT model to fill the gaps, i.e translate all available ASR data. This allows us to increase the amount of training data for SLT significantly, especially for

Pair/Model	TF	+Rel	+MCR	+16L	+ESB	+DSF	+ESB	+DSF2
es-en	33.48	33.98	34.94	34.93	35.16	35.88	**36.14**	35.83
en-es	30.87	31.34	31.88	31.72	32.76	33.42	**33.97**	33.56
es-fr	40.65	41.40	41.19	41.26	42.06	42.87	**43.57**	43.12
fr-es	38.48	38.59	38.98	38.85	39.87	40.82	**41.09**	40.88
es-it	28.82	29.07	30.24	31.29	31.27	32.50	**33.80**	32.93
it-es	34.74	35.27	35.25	35.31	36.58	38.41	**39.01**	38.50
es-pt	43.04	43.40	43.65	43.53	44.30	44.96	**45.40**	45.03
pt-es	46.95	47.01	46.63	46.59	47.70	48.74	**48.95**	48.41
fr-en	38.29	38.62	39.64	39.53	40.32	41.09	**41.65**	40.93
en-fr	39.88	40.47	40.85	41.18	41.51	42.40	**43.17**	42.14
fr-pt	40.61	41.31	41.71	42.52	42.50	43.94	**44.25**	43.52
pt-fr	46.14	46.42	46.57	47.02	47.76	48.90	**49.66**	48.76
fr-pt	37.67	38.49	38.73	39.81	39.57	40.23	**40.55**	39.52
pt-fr	34.60	34.53	35.07	35.43	35.58	36.59	**37.05**	36.51
avg	38.16	38.56	38.95	39.21	39.78	40.76	41.3	40.68
		+0.4	+0.29	+0.26	+0.57	+0.98	+0.54	-0.62

Table 4: IWSLT 2021 machine translation progressive results. The features including Relative Attention (REL), Macaron FFN (MCR), 16 layer-deep (16L), ensembling (ESB) and diversification (DSF) are additive from left to right, starting from the base model. The last row shows the improvement compared to the previous increment.

languages such as Italian and French.

Architecture wise, we only used Transformers for SLT, that followed the same training procedure with ASR due to the fact that the encoders are transferred from the Transformer ASR models.

The results are shown in Table 5. Unfortunately the results without ASR pre-training are not available because training was unstable and likely to diverge in such harsh data condition. It is not unexpected that the end-to-end model (E2E) trained with only the initially limited amount of data falls behind the performance of the cascade models. With distillation from machine translation, the performance is largely boosted to be on par with the cascade. The 0.2 differential in average mostly comes from Portuguese-Spanish, Italian-English and Italian-Spanish.

Compared with pre-distillation, a lot of language-pairs enjoyed a significant improvement of up to 26 BLEU points, such as the sample Italian audio inputs, thanks to the distillation models changing zero-shot to supervised settings. The supervised language pair that was mostly improved is Spanish-French (12 BLEU points).

Finally, in this particular SLT setup, we found that it is useful to ensemble cascade and SLT models in a multi-modal manner. In the literature, it has been observed that each approach has its own strength. While the components of the cascade can be easily tuned individually because ASR and MT have lower mapping complexity than SLT, the end-to-end models can avoid error-propagation that plagues cascade systems. An ensemble suggests that we can combine the strengths of two approach, yet only available in certain experimental settings that leaves *audio segmentation* out of the scope. Here the ensemble is done by simply using the same bpe vocabulary for the MT and SLT models, and average the output probabilities of the MT and SLT models for every timestep. The result showed that this intuition can help improve the results further.

7 Final submission

Our final submissions include an ensemble of E2E and Cascade as primary, with the E2E model served as the contrastive. The official results are shown in Table 6.

In the final results, we can see that the ensemble quality depends on the ASR performance, which can be seen in test sets with Spanish audio and French audio. At the relatively low error rate, combining two approaches provides a significant boost to the translation quality. However, for French samples the deterioration of the cascade makes the combination worse than the sole end-to-end solu-

Model Pair	Cascade	E2E	+Syn	+ESB
es-en	30.44	25.58	30.27	**31.02**
es-fr	31.64	18.81	31.32	**32.25**
es-it	26.07	22.94	26.22	**26.21**
es-pt	39.33	34.73	39.53	**40.04**
fr-en	35.41	29.73	35.19	**36.06**
fr-es	37.71	30.13	38.48	**38.96**
fr-pt	38.21	30.98	37.97	**38.44**
pt-en	33.63	28.16	33.25	**34.15**
pt-es	37.53	25.55	38.41	**38.43**
it-en	24.28	5.37	24.92	**25.29**
it-es	32.29	7.20	33.67	**33.90**
avg	33.32	23.56	33.56	**34.06**

Table 5: End-to-end speech translation results on progressive testsets.

SLT Pair	Ensemble	E2E
es-en	39.3	38.9
es-fr	32.4	31.4
es-it	32.3	31.4
es-pt	46.6	46.7
fr-en	27.1	28.5
fr-es	29.2	29.7
fr-pt	28.8	28.7
pt-en	30.7	30.2
pt-es	37.3	37.1
it-en	26.5	25.8
it-es	32.4	33.0
ASR		
es	10.0	-
fr	26.5	-
it	15.5	-
pt	22.1	-

Table 6: Official IWSLT 2021 Speech recognition and translation results.

tion.

This experiment shows that error propagation is a serious problem and end-to-end SLT systems can be more robust than cascades with sufficient data and training efficiency improvement.

The evaluation also suggests us to investigate into zero-shot translation for multilingual SLT, which is extremely difficult because of the modality difference between the source and target sequences.

Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n°825460 and the Federal Ministry of Education and Research/DLR Projektträger Bereich Gesundheit (Germany) under grant agreement n° 01EF1803B.

References

Ebrahim Ansari, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, et al. 2020. Findings of the iwslt 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pages 308–327.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11.

Yarin Gal and Zoubin Ghahramani. 2016. Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *4th International Conference on Learning Representations (ICLR) workshop track*.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective strategies in zero-shot neural machine translation. *arXiv preprint arXiv:1711.07893*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE.

Niehues Jan, Roldano Cattoni, Stuker Sebastian, Matteo Negri, Marco Turchi, Salesky Elizabeth, Sanabria Ramon, Barrault Loic, Specia Lucia, and Marcello Federico. 2019. The iwslt 2019 evaluation campaign. In *16th International Workshop on Spoken Language Translation 2019*.

Niehues Jan, Roldano Cattoni, Stuker Sebastian, Marco Turchi, Mauro Cettolo, and Marcello Federico. 2018. The iwslt 2018 evaluation campaign. In *15th International Workshop on Spoken Language Translation 2018*.

M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viegas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2019a. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. *arXiv preprint arXiv:1910.13296*.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2019b. Data diversification: A simple strategy for neural machine translation. *arXiv preprint arXiv:1911.01986*.

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. *arXiv preprint arXiv:1904.08779*.

Ngoc-Quan Pham, Thanh-Le Ha, Tuan-Nam Nguyen, Thai-Son Nguyen, Elizabeth Salesky, Sebastian Stüker, Jan Niehues, and Alex Waibel. 2020. Relative Positional Encoding for Speech Recognition and Direct Translation. In *Proc. Interspeech 2020*, pages 31–35.

Ngoc-Quan Pham, Thai-Son Nguyen, Thanh-Le Ha, Juan Hussain, Felix Schneider, Jan Niehues, Sebastian Stüker, and Alexander Waibel. 2019a. The iwslt 2019 kit speech translation system. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019b. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, 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. *arXiv preprint arXiv:2102.01757*.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.