BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Aligning Pre-Trained Models for Spoken Language Translation

Bc. Šimon Sedláček*
Supervisor: Santosh Kesiraju Ph.D.

Abstract
Speech translation (ST) systems most commonly adopt either a cascade or an end-to-end (E2E) approach. While cascade systems do not require any tuning on ST data, they can suffer from greater model latency and error accumulation. On the other hand, training an E2E ST system can be resource intensive, both data and computation-wise. In this work, we investigate the possibilities of leveraging models pre-trained models for source language ASR and source-to-target language MT, connecting them with a small connector module (Q-Former) to solve the ST task. While keeping the speech encoder and MT decoder frozen, the connector module is trained to bridge gap between the speech and text modalities, transforming the ASR encoder embeddings into the space of the MT encoder text embeddings. We train and evaluate our models on the How2 English to Portuguese ST dataset. In our experiments, we find that all aligned models outperform the cascade ST model baseline. Additionally, while keeping the size of the connector module constant, increasing the size and capability of the ASR encoder and MT decoder improves the translation results, outperforming even the in-domain encoder-decoder fine-tuned baseline ST system, while having to tune fewer parameters.

*xsedla1h@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Speech translation (ST) is the task of mapping an input speech utterance in a certain *source language* to its corresponding text translation in a given *target language*. The two main ST system architectures (see Figure 1) are *cascade* and *end-to-end* (E2E), typically implemented using various deep learning models.

With modern speech recognition (ASR) and MT systems getting more and more reliable, cascade approaches often yield state-of-the-art performance. On the other hand, such models tend to have bigger latency and can suffer from error accumulation, as the models operate disjointly. E2E ST systems mitigate these drawbacks by establishing a differentiable path from the source audio to the output translation[1]. However, they additionally require ST data for training and fine-tuning of the whole model, in contrast to cascade approaches.

A potential middle ground could be to leverage powerful off-the-shelf pre-trained ASR and MT models, freeze them, and align their representation spaces with a small connector module, so that the final model solves the ST task. This is inspired by other approaches that have recently been adopted in multiple domains, such as vision-language[2, 3] and speech recognition[4, 5].

2. Model Alignment

The proposed alignment framework is shown in Figure 2. First, we choose a pre-trained source language speech encoder and a source-to-target language MT model. These two models can have completely different sizes, hidden dimensions, numbers of layers. Then, the models are *frozen* and connected via one of the two connector/alignment model variants.

Variant A is the *Q-Former*[3, 4]. The Q-Former is a transformer decoder model that uses a sequence of trainable *query* vectors as its input. These queries then interact with the *ASR encoder output* hidden representations via cross-attention, gradually extracting more information about the source utterance.

The number of queries is a hyperparameter – the Q-Former maps variable-length speech embedding sequences into the fixed-length query space.

Variant B has the architecture of a plain *transformer encoder*. The input speech embeddings are first subsampled to 1/4 length with a 1D conv. subsampler, and then used as input to the encoder, doing away with the fixed-length mapping problem of the Q-Former.

The output of the connector is subsequently passed to the MT decoder via cross-attention, generating the translations and allowing to optimize for the language modeling objective.

3. Data and evaluation

All models in this work were trained and evaluated on the How2 dataset[6]. The How2 dataset is a multimodal corpus of English instructional videos, and their respective transcriptions. It has a smaller 300-hour audio subset with Portuguese translations, details for which are shown in Table 1.

Models are evaluated using the standard BLEU[7] metric for translation systems. For ASR performance, the word-error-rate (WER) metric is used. All systems are evaluated on both the `val` and `dev5` sets.

4. Base ASR and MT models

We conduct our experiments with two base ASR and two base MT models. For the ASR models (see Table 2), we first train a 38.5M parameter hybrid CTC/attention[8] *E-Branchformer*[9] *base* model on the How2 dataset. This model provides an in-domain ASR reference point, as it achieves similar WER on How2 test sets as models from ESPnet-ST[10].

The second ASR system is a similar E-Branchformer model trained with *decoder-centric regularization*, a novel ASR training method developed by Ing. Alexander Polok and BUT@Speech. The method has not been published yet, however, the model is freely available on HuggingFace[1]. This model is out-of-domain on How2, however, it has been trained 6000 hours of English data, providing an excellent baseline for a good off-the-shelf ASR model. In Table 2, it is referred to as E-Branchformer medium.

For the MT systems, a small in-domain MT model based on the MarianMT[11] was trained on How2. The other MT system is an off-the-shelf out-of-domain T5 encoder-decoder model[12], pre-trained on Portuguese language and then fine-tuned for `en->pt` translation. The performances of both systems on How2 are shown in Table 3.

5. Experiments

Using the E-Branchformer base and MarianMT models, two baseline ST models are constructed. The first is a cascade system with an additional *truecaser* model to overcome the discrepancy between the vocabularies of the ASR and MT models. The second one is a conventional E2E ST system constructed from the ASR encoder and MT decoder of these two models and then fully fine-tuned (FT) on How2.

For the alignment experiments, the pre-trained MarianMT encoder weights were used to initialize both variant A (Q-Former with 128 queries) and variant B (conv. + MT encoder) of the alignment framework. The connector module therefore has 6 transformer layers with 4 attention heads and hidden size of 256. All aligned models are trained using HuggingFace Transformers[2] on the BUT FIT SGE cluster. Models are trained for 70 epochs maximum with a batch size of 128 and learning rate of 2×10^{-4}.

The results in Table 4 clearly show that while only training the connector module does not yield as good of a performance as the fully FT baseline system, the results are still reasonable in comparison to the cascade system. A clear trend is that better ASR and MT models (and more so ASR than MT) lead to better aligned system performance, with the best one achieving 46.8 BLEU on the `val` set.

Additionally, it seems that for the ST task, the alignment variant B seems to perform marginally better, possibly due to losing information to the fixed-length representation mapping of the Q-Former.

Lastly, even though the T5 model only achieved 40 BLEU on the `val` set, it still provides a performance boost in the aligned setting, suggesting that the Q-Former can serve as domain adapter as well.

6. Conclusions

The experiments conducted in this work show that aligning frozen ASR encoders with MT decoders is a viable and generic approach to training speech translation systems, yielding good ST performance even in comparison to conventional methods, while requiring fewer parameters to train. Going further, we aim to analyze the strengths and drawbacks of the two alignment models, and develop a pre-training approach, which could improve the ST performance in lower resource scenarios.

[1] huggingface.co/BUT-FIT/EBranchRegulaFormer-medium

[2] https://huggingface.co/

References

[1] Santosh Kesiraju, Marek Sarvaš, Tomáš Pavlíček, Cécile Macaire, and Alejandro Ciuba. Strategies for improving low resource speech to text translation relying on pre-trained asr models. In *INTERSPEECH 2023*. ISCA, August 2023.

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[4] Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Connecting speech encoder and large language model for asr, 2023.

[5] Yukiya Hono, Koh Mitsuda, Tianyu Zhao, Kentaro Mitsui, Toshiaki Wakatsuki, and Kei Sawada. An integration of pre-trained speech and language models for end-to-end speech recognition, 2023.

[6] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding, 2018.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[8] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In *Proc. Interspeech 2019*, pages 1408–1412, 2019.

[9] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. E-branchformer: Branchformer with enhanced merging for speech recognition, 2022.

[10] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. Espnet-st: All-in-one speech translation toolkit, 2020.

[11] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in c++, 2018.

[12] Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. Lite training strategies for Portuguese-English and English-Portuguese translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 833–840, Online, November 2020. Association for Computational Linguistics.