TACKLING DATA SCARCITY IN SPEECH TRANSLATION USING ZERO-SHOT MULTILINGUAL MACHINE TRANSLATION TECHNIQUES

Tu Anh Dinh, Danni Liu, Jan Niehues

Department of Data Science and Knowledge Engineering
Maastricht University, The Netherlands

ABSTRACT

Recently, end-to-end speech translation (ST) has gained significant attention as it avoids error propagation. However, the approach suffers from data scarcity. It heavily depends on direct ST data and is less efficient in making use of speech transcription and text translation data, which is often more easily available. In the related field of multilingual text translation, several techniques have been proposed for zero-shot translation. A main idea is to increase the similarity of semantically similar sentences in different languages. We investigate whether these ideas can be applied to speech translation, by building ST models trained on speech transcription and text translation data. We investigate the effects of data augmentation and auxiliary loss function. The techniques were successfully applied to few-shot ST using limited ST data, with improvements of up to +12.9 BLEU points compared to direct end-to-end ST and +3.1 BLEU points compared to ST models fine-tuned from ASR model.

***Index Terms*—** speech translation, zero-shot, few-shot, machine translation, multi-task

1. INTRODUCTION

Speech Translation (ST) is the task of translating speech in a source language into text in a target language. Traditional ST approaches include two cascaded steps: Automatic Speech Recognition (ASR) and Machine Translation (MT). These approaches are prone to errors propagated from the ASR step to the MT step [1]. Due to that, end-to-end ST has recently been gaining more interest [2], [3], [4], [5]. End-to-end ST translates source-language speech directly into target-language text, thus less prone to error propagation. However, end-to-end ST performance heavily depends on direct ST data, which can be difficult to obtain. ASR and MT data, on the other hand, is often more easily available [6].

To tackle data scarcity, it is useful to make use of ASR and MT data for end-to-end ST models. In order to achieve that, we explore techniques from zero-shot multilingual text translation and apply them to speech translation in this paper. Zero-shot, in the context of translation models, is an approach that enables translating a pair of languages that is unseen during training [6]. We adapt this idea to speech translation by training a single multi-task model on ASR and MT data to perform the ST task in an end-to-end manner. We focus on a few-shot scenario, when a limited amount of ST data is available. In this scenario, we fine-tune our multi-task model with a small amount of ST data before performing ST task.

Motivated by the findings in multilingual text translation, we address two challenges to increase the efficiency of exploiting the ASR and MT data for ST task. First, we address the representation of semantically similar input using different modalities. Second, we address the model's ability to control the output language. We tackle the first challenge by using the Transformer architecture with shared encoder, and use an auxiliary loss function to minimize the difference in text and audio representation [7]. For the second challenge, we propose a data augmentation approach to control the output language. We find these approaches to be particularly useful in the few-shot scenario. Our multi-task few-shot models outperform direct ST models trained on the same limited ST data from scratch by up to +12.9 BLEU points. Our models also outperform ST models fine-tuned from an ASR model using the same ST data by up to +3.1 BLEU points. This proves that our models have successfully made use of ASR and MT data to improve ST performance. The implementation is at `https://github.com/TuAnh23/MultiModalST`.

2. RELATED WORK

Different approaches for ST have been explored over the decades, as summarized in [6]. End-to-end ST, which is expected to overcome the error-propagation issue of the traditional cascaded approach, has recently been the method of interest. Examples of end-to-end ST include direct ST models trained on ST data from scratch [2], models pre-trained on ASR task and fine-tuned on ST task [3], models trained on ST data generated by augmenting ASR or MT data [4] and models co-trained on MT and ST task [5]. A major challenge of end-to-end ST is the lack of direct ST data for training.

Zero-shot translation has been shown to work for multilingual MT. In [8], zero-shot multilingual MT is enabled

This paper was done in collaboration with Mediaan, Heerlen.

by adding a language token to the beginning of the input sequence to indicate the required target language. Several studies have been done to improve the quality of zero-shot multilingual MT. Approaches to encourage a source-language-independent representation are proposed in [7], such as using a fixed size encoder for different languages. In [9], a language-independent representation is encouraged by disentangling positional information of the input and output tokens.

Inspired by the ability of zero-shot for multilingual MT, we study the applicability of similar approaches on ST tasks, by building an ST multi-task model trained on ASR and MT tasks, which data is often more easily available than ST.

3. MULTI-TASK MODEL

We propose a multi-task model as illustrated in Fig. 1 to reduce the need for direct ST training data. The model is trained simultaneously on ASR task and MT task. We consider the few-shot scenario when a limited amount of ST data is available. We build few-shot model by fine-tuning the above multi-task models with a small amount of ST data before performing the ST task. We also attempt on zero-shot ST, using the multi-task model directly for ST task without fine-tuning. The requirement is that the model represents *SRC audio* and *SRC text* in a similar way so that it can leverage the ASR and MT tasks learnt during training to perform the ST task.

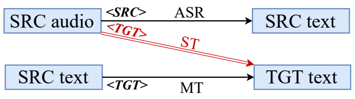

Fig. 1. Proposed multi-task model. Single-line arrows are training directions. Double-line arrow is inference direction. Tags in the brackets are target-language tokens.

We build upon the deep Transformer [10] for speech proposed in [11]. To enable encoding audio and text jointly, we share the model parameters between the two modalities. As we expect the bottom layers of the audio encoder to extract low-level acoustic features, we assign more layers to the audio encoder and share its top layers with the text encoder. This shared encoder is crucial to make audio and text representations similar. The overall structure is shown in Fig. 2. To indicate output language, we apply the same method as for zero-shot multilingual MT. We add target-language tokens to the beginning of input sequences and concatenate the target language embeddings to every decoder input to enforce the model outputting the language of interest [7]. Target-language tokens are shown in the brackets in Fig. 1.

The proposed model introduces a challenge, where it outputs the wrong language for ST task. It outputs *SRC text* for *SRC audio* input, even when the target token is <*TGT*>. The reason is that in the training data, audio input always has *SRC* output and text input always has *TGT* output, thus the models decide on output language based on input modality instead of the target-language tokens. This boils down to two issues: the difference between modalities (text and audio), and the reliance of the model on the target-language tokens. We propose approaches to tackle these issues as follows.

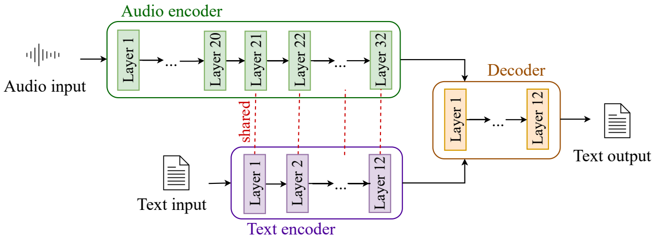

Fig. 2. Overall structure of multi-task models.

3.1. Cross-modality knowledge sharing

A prerequisite to knowledge sharing across the ASR and MT tasks is common representations across input modalities, as shown in Fig. 3. We encourage modality-independent representation of the data by using an auxiliary loss function that minimizes the difference between encoder output of audio and text. The metric for the difference is the squared error of mean-pool over time [7]. That is, given a pair of aligned text sentence X and audio sentence Y, the auxiliary loss is: $[mean\_pool(Encoder(X)) - mean\_pool(Encoder(Y))]^2$.

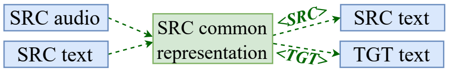

Fig. 3. Motivation for cross-modality knowledge sharing.

3.2. Controlling output language

A challenge for the model in Fig. 1 is to control the output language. As mentioned above, each modality has only one target language during training, leading to the model deciding on output language based on input modality, instead of the specified target-language tokens as desired.

To have stronger control on the output language, we propose data augmentation to increase the model's reliance on the provided target tokens. The idea is to include artificial data so that each modality has more than one output language during training. We expect this to force the model to rely on the target tokens to decide which language to output. Data augmentation would avoid the need for another real dataset.

We introduce an artificial language by reversing *SRC* sentences character-wise, denoted as *SRC-R*. By reversing character-wise, the new language's vocabulary would not conflate with the original language. The training data is shown in Fig. 4, along with the annotations. By adding (1) and (2), *SRC audio* and *SRC text* have two target languages during training. By adding (3) and (4), the model can learn to switch between outputting *SRC* and *TGT* languages.

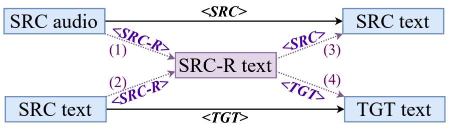

Fig. 4. Augmented training data. Solid-line arrows are main directions. Dotted-line arrows are artificial directions.

4. EXPERIMENTAL SETUP

The experiments goal is to build and evaluate ST models to translate English (*EN*) audio to German (*DE*) text.

4.1. Dataset and preprocessing

We use the CoVoST 2 speech translation corpus [12]. The data used is English to German pair, with 289K training samples (430 hours of speech), 15K validation samples (26 hours of speech) and 15K testing samples (25 hours of speech).

For audio data, we extract and normalize 40-dimensional log scale mel filterbank concatenated with its delta coefficient to use as input features. For text data, we remove the double quotes at the beginning and the end of the sentences, and use SentencePiece [13] without pre-tokenization and pre-normalization to build subword-based vocabularies.

4.2. Model configurations

Our models use the Transformer architecture with attention-based encoder and decoder [10, 11]. We adapt the hyperparameters choices in [14] to our multi-modal setting: 32 audio encoder layers, 12 text encoder layers, 12 decoder layers; the 12 text encoder layers are shared with the top 12 audio encoder layers. The other configurations are the same as [14].

When training multiple datasets in one model (e.g., ASR and MT), the batches from each dataset are ordered alternatively. The models are trained for 64 epochs, except for the fine-tuned models, where we stop training as soon as the validation loss stops reducing. The checkpointed model with the lowest validation loss is used for final evaluation.

For ASR tasks, the models are evaluated using lowercased, tokenized, punctuation-removed Word Error Rate (WER) calculated using VizSeq [15]. For MT and ST tasks, the metric used is the detokenized, case-sensitive BLEU score, calculated using sacreBLEU [16].

5. RESULTS AND DISCUSSION

The transcription and translation quality of our few-shot models is shown in Table 1. In this few-shot scenario, we only use a limited amount (10% and 25%) of ST data. The amount of ASR and MT data used is 100%.

We present three baselines using the same limited ST data: (1) direct end-to-end ST model trained on ST data from scratch, (2) ST model fine-tuned from a pre-trained ASR model (indirectly making use of ASR data) and (3) ST model fine-tuned from plain multi-task model (indirectly making use of ASR and MT data). As shown in Row 1 of Table 1, the direct end-to-end ST model fails to perform the ST task given the extremely limited ST data. In contrast, the ST model making use of ASR data is able to perform ST task (Row 2). Our plain few-shot model, additionally making use of the MT data, has better ST performance, although the gain from the MT data is quite limited (Row 3).

In Row 4-7, we denote our approaches with P + $\{$*approach name*$\}$. We observe that our approaches improve the ST performance of the few-shot multi-task model. Data augmentation (Row 6) gives higher improvement than auxiliary loss (Row 4). The combination of both approaches (Row 7) gives the best performance, with the improvement of up to +12.9 BLEU points compared to the direct end-to-end baseline, +3.1 BLEU points compared to the fine-tuned ASR baseline and +1.7 compared to the plain multi-task baseline. We also observe that, as we increase the amount of ST data (25% ST data instead of 10%), the gains from our approaches generally become less significant. This suggests that our approaches are particularly effective in low-resource scenarios.

A challenge of data augmentation is that the model cannot learn ASR task well, leading to poor few-shot ST performance (Row 5), due to the high number of tasks in augmented data. Therefore, instead of training the model on augmented data from scratch, we first train it on the original data (ASR and MT), and then fine-tune with augmented data. In this way, the issue no longer persists (Row 6 and 7). It seems important for the model to first learn the easier task (ASR) and then ST task. This can be viewed as a type of curriculum learning.

We also experimented on a direct ST model trained on 100% of the ST data for comparison. The BLEU score of this model is 14.9. The BLEU score of our best few-shot model is 13.7. Our best few-shot model uses only 25% of the ST data for training, yet only falls short by 1.2 BLEU points compared to this direct model using 100% of the ST data.

We also attempted on zero-shot using no ST data. This remains challenging, since the BLEU scores are under 1. However, the models using data augmentation no longer output all wrong language (as discussed in Section 3), proving the effectiveness of data augmentation.

6. ANALYSIS

6.1. Cross-modal similarity and translation quality

We study the similarity of representations between modalities under different approaches and analyze them in relation to

Table 1. Performance summary of different model types.

No.	Model type	ASR (WER ↓) (%)	MT (BLEU ↑)	Few-shot ST (BLEU ↑) [a] 10% ST data	Few-shot ST (BLEU ↑) [b] 25% ST data
1	Single ST task / Direct end-to-end ST	-	-	0.5	0.8
2	Single ASR task	25.8	-	8.4	10.9
3	Plain multi-task (P)	28.4	32.8	9.8	12.4
4	P + auxiliary loss	26.9	32.5	10.6 (**+10.1** | **+2.2** | **+0.8**)	13.2 (**+12.4** | **+2.3** | **+0.8**)
5	P + augmented data	47.6	31.6	5.4 (**+4.9** | -3.0 | -4.4)	6.4 (**+5.6** | -4.5 | -6.0)
6	P + augmented data (fine-tune from plain)	27.6	32.2	11.5 (**+11.0** | **+3.1** | **+1.7**)	13.5 (**+12.7** | **+2.6** | **+1.1**)
7	P + augmented data + auxiliary loss (fine-tune from plain)	27.7	32.3	11.5 (**+11.0** | **+3.1** | **+1.7**)	13.7 (**+12.9** | **+2.8** | **+1.3**)

[a,b] Additionally reporting the differences compared to the three baselines in Row 1, Row 2 and Row 3, respectively.

few-shot ST performance. We measure representational similarity between text and audio encoder output using Singular Vector Canonical Correlation Analysis (SVCCA) [17], where higher scores indicate higher similarity.

The results are summarized in Fig. 5, associating the similarity scores with few-shot translation quality. Observe that all proposed approaches increase text-audio similarity. The results also agree with our hypothesis: more text-audio similarity means higher BLEU score, i.e., better ST performance.

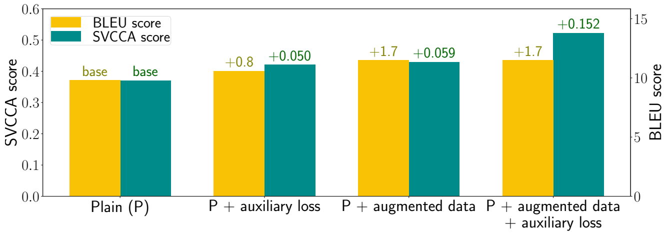

Fig. 5. SVCCA analysis along with translation quality. Numbers on the bars are comparision to the plain model.

6.2. Cross-modal similarity at token and frame level

In addition to the SVCCA score, which measures text-audio similarity on a sentence level, we use another method to quantify text-audio similarity on a token level. We train a classifier to predict input modalities based on encoder outputs. Better classification performance indicates higher dissimilarity between the modalities [9]. Since the number of audio frames is significantly higher than the number of text tokens, we consider the True Positive Rate (TPR, proportion of audio tokens identified correctly) and the True Negative Rate (TNR, proportion of text tokens identified correctly) instead of the overall accuracy. The result is shown in Table 2. We observe that for all models that do not use auxiliary loss, both the TPR and TNR are over 99.9%, meaning that text and audio encoder output tokens are very distinguishable. On the other hand, for all models using auxiliary loss, the TPR is over 99.9% and the TNR is under 10%. This means that the classifier has a poor performance where it predicts most of the tokens as audio, which indicates high similarity between text and audio encoder output tokens. Thus, we conclude that using auxiliary loss indeed increases text-audio similarity on a token level.

Table 2. Performance of token modality classifier on the encoder outputs of different multi-task models.

Multi-task model	TPR (%)	TNR (%)
Plain multi-task (P)	99.99	99.89
P + augmented data	99.99	99.99
P + auxiliary loss	99.71	**9.77**
P + augmented data + auxiliary loss	99.99	**.82**

7. CONCLUSIONS

In this paper, we study how to alleviate the data scarcity problem of end-to-end ST by utilizing zero-shot multilingual MT techniques. Based on a multi-task model for ASR and MT, we propose approaches to (1) encourage knowledge sharing between text and audio modalities and (2) enforce stronger control of the output language. Our approaches successfully make use of ASR and MT data in the few-shot scenario, and improve the ST performance by up to +12.9 BLEU points compared to direct end-to-end ST models and +3.1 BLEU points compared to ST models fine-tuned from ASR models.

8. REFERENCES

[1] Nicholas Ruiz and Marcello Federico, "Assessing the impact of speech recognition errors on machine translation quality," in *11th Conference of the Association for Machine Translation in the Americas (AMTA), Vancouver, BC, Canada*, 2014.

[2] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Low-resource speech-to-text translation," *Annual Conference of the International Speech Communication Association (InterSpeech)*, 2018.

[3] Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 58–68.

[4] Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7180–7184.

[5] Yun Tang, Juan Pino, Changhan Wang, Xutai Ma, and Dmitriy Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6209–6213.

[6] Matthias Sperber and Matthias Paulik, "Speech translation and the end-to-end promise: Taking stock of where we are," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, pp. 7409–7421.

[7] Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel, "Improving zero-shot translation with language-independent constraints," in *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Aug. 2019, pp. 13–23.

[8] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[9] Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li, "Improving zero-shot translation by disentangling positional information," *arXiv preprint arXiv:2012.15127*, 2020.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[11] Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[12] Changhan Wang, Anne Wu, and Juan Pino, "Covost 2: A massively multilingual speech-to-text translation corpus," *arXiv preprint arXiv:2007.10310*, 2020.

[13] Taku Kudo and John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, Nov. 2018, pp. 66–71, Association for Computational Linguistics.

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

[15] Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu, "Vizseq: A visual analysis toolkit for text generation tasks," in *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2019.

[16] Matt Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels, Oct. 2018, pp. 186–191, Association for Computational Linguistics.

[17] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 6076–6085. Curran Associates, Inc., 2017.