VUS at IWSLT 2021: A Finetuned Pipeline for Offline Speech Translation

Yong Rae Jo
Voithru Inc.

Young Ki Moon
Voithru Inc.

Minji Jung
Voithru Inc.

Jungyoon Choi
Voithru Inc.

Jihyung Moon
Upstage

Won Ik Cho
Seoul National University

{yongrae.jo, minji.jung, jungyoon.choi}@voithru.com
ykmoon0814@gmail.com, jihyung.moon@upstage.ai
tsatsuki@snu.ac.kr

Abstract

In this technical report, we describe the *fine-tuned*[1] ASR-MT pipeline used for the IWSLT shared task. We remove less useful speech samples by checking WER with an ASR model, and further train a wav2vec and Transformers-based ASR module based on the filtered data. In addition, we cleanse the errata that can interfere with the machine translation process and use it for Transformer-based MT module training. Finally, in the actual inference phase, we use a sentence boundary detection model trained with constrained data to properly merge fragment ASR outputs into full sentences. The merged sentences are post-processed using part of speech. The final result is yielded by the trained MT module. The performance using the dev set displays BLEU 20.37, and this model records the performance of BLEU 20.9 with the test set.

1 Introduction

Offline speech translation is a task that infers the text of a target language by using speech as input. A pipeline system is used as a representative method, which converts source speech into the source text via automatic speech recognition and machine translates it. Recently, many speech corpora have been disclosed, and studies are being conducted on an end-to-end method, namely directly decoding speech input into the text of a target language (Bérard et al., 2016, 2018).

In this IWSLT shared offline task, we implement an English-German speech translation system in a pipeline format. The advantage of pipeline architecture is that it can explain whether the given speech translation is challenging in view of the acoustic domain or the translation perspective, considering the whole process of converting source speech to the target text. This makes it easier for us to discern difficult or erroneous parts in speech and text processing.

In general, a limitation of a pipeline system compared to an end-to-end system is that the quality of the final result is largely influenced by the intermediate text representation, which is usually obtained in an explicit format (Liu et al., 2020). Therefore, we primarily remove training samples that can lower the ASR performance, following the method used in Potapczyk and Przybysz (2020). Thereafter, based on the trained ASR module, the output of test speech samples is transformed into the text and fed to the machine translation system to produce a final output. In this process, we conduct post-processing to obtain an accurate sentence-level output, such as setting the sentence boundary between the fragment texts and re-aggregating some wrongly merged sentences.

The performance is checked mainly with BLEU score (Papineni et al., 2002). Through the system construction, we obtained a BLEU score of 20.9 in en-de speech translation. In detail, the performance of the ASR module reaches WER 28.3% based on 2015 test set, and the MT module records a BLEU score of 32.2 based on the WMT dataset (Barrault et al., 2020). In addition, we have observed that various pre- and post-processings lead to meaningful performance gains.

In this paper, we first skim the related works on speech translation, automatic speech recognition, and machine translation, focusing on the publicly available datasets. Then we describe how we obtained the ASR and MT module used for the campaign. Next, we demonstrate how we finally reach the translation for the dev and test set, along with some pre- and post-processing techniques. The results are provided with the analysis.

[1]We use 'fine-tuned' to describe that our approach is not fully end-to-end but incorporates a well-organized set of strategies to reach better performance. It does not denote the wav2vec-transformer ASR module either.

Proceedings of the 18th International Conference on Spoken Language Translation, pages 120–124
Bangkok, Thailand (Online), August 5–6, 2021.

2 Related Work

Various datasets exist for speech translation using English as the source language, being utilized in the training and evaluation in a wide range of studies. The representative one is MuST-C (Di Gangi et al., 2019), which provides English speech of TED talks, its transcript, and the translation to other Indo-European languages, including German, where we exploit en-de in this study. In addition, CoVoST enables multilingual speech translation based on Common Voice (CV) data (Wang et al., 2020), of which the Wikipedia articles are the source text. Europarl-ST (Koehn, 2005) also provides various translations, for the debates in European Parliament.

Data used for speech translation can also be used for automatic speech recognition and machine translation, but there are also corpora built for ASR and MT only, on a large scale. Librispeech (Panayotov et al., 2015), which is used for evaluation of ASR models, is the most famous example, and TedLium is also the case[2]. They consist of the speech of the source language (English) and Latin alphabet-based transcription. In contrast, since only text data is used in MT, the scale is much larger. Typically used sources are WMT datasets (Bojar et al., 2016, 2018; Barrault et al., 2020) and Open subtitles. [3] All of the above datasets can be usefully used in speech translation, so they have been actively utilized in the previous IWSLT campaigns (Niehues et al.).

3 Model

We chose the cascading scheme to leverage the high performance of ASR and MT modules. Thus, we exploit a large variety of corpora mentioned above to train each module.

3.1 Automatic Speech Recognition

We train the ASR module using Librispeech and MuST-C. The pretrained wav2vec 2.0 base model was used for embedding (Baevski et al., 2020), and the training was conducted with a Transformer (Vaswani et al., 2017) decoder part augmented on the output layer of the wav2vec module, with character as vocab. In this process, we performed two preprocessing for the source corpus.

- **Script normalization:** In the sentences containing laughter and applause tag, the expressions that might deter ASR performance were removed.

- **Filtering out erroneous scripts:** Following SRPOL's approach (Potapczyk and Przybysz, 2020), we performed the filtering of audio files based on bad WER. In this process, sentences showing WER below 75% were removed, assuming as if there were some flaws in the acoustic level or some errors in the script.

Using the cleansed corpus created through the above process, we conducted the training for 80,000 steps using 8 RTX 3090 devices. The optimization was done with adam, learning rate 1e-5, and dropout 0.1. As a result of utilizing the evaluation set 2015 test set, we obtained an ASR module that displays the WER of 28.3%.

3.2 Machine Translation

We trained the MT module using the WMT 20 en-de news task dataset and Transformer architecture.

For English, the script was normalized, and for German, the cased text was used. Vocabulary was constructed in consideration of both English and German, using subword tokenization (Sennrich et al., 2016). Some preprocessings were performed as follows:

- **Language identification:** We conduct language identification to remove the instances where the source and the target language do not match the language of interest (en, de). This refers to Lui and Baldwin (2011, 2012); Heafield et al. (2015).

- **Filter by length:** We filter out the sentences where the length of the source and the target sentence displays more than 50% of difference.

- **Written-to-spoken text conversion:** We first transform the source text into the format of speech transcript, namely lowercasing the text and removing all punctuation marks. Then we expand common abbreviations, especially for measurement units, by converting numbers, dates, and other entities expressed with digits into their spoken form. The overall scheme follows Bahar et al. (2020).

[2] https://www.openslr.org/7/
[3] https://www.opensubtitles.org/

Using the cleansed WMT script, we conducted the training for 300,000 steps, using 8 RTX 3090 devices. The optimization was done with adam, with FFN decoder 8,192 and dropout 0.1. With WMT20 dev set, we obtained an MT module that shows the BLEU of 32.3.

4 Inference

We infer the final output with the speech instances of the dev set using the trained ASR and MT modules. After the inference, we submit the inference of the test set using the model that yields the best results with the dev set.

In the inference process of the dev and test set, a proper sentence split is additionally required. For the dev and test set, we separated the utterances from silence using the given segmentation information. The segmented audio files were transcribed with the ASR module.

In the post-processing of the transcribed speech, we use the following strategies.

- **DeepSegment**: We merge the output of the ASR module using publicly available DeepSegment recipe[4] based on bidirectional long short term memory and conditional random field (BiLSTM-CRF) (Huang et al., 2015). At this time, the BiLSTM-CRF model is trained using 1 RTX TITAN. Here, no information other than the training corpus is used for the training, and the usage of NLTK in featurization does not violate the constrained condition.

- **Sentence concatenation**: We compensate for probable segmentation errors by using part-of-speech (POS) information. We selected POS tags that are rarely placed in sentence-first and sentence-final from 46 tags of NLTK POS tagger (Loper and Bird, 2002). In detail, we set two cases of `PROHIBIT_AS_FIRST` and `PROHIBIT_AS_FINAL` as follows:

  - `PROHIBIT_AS_FIRST`: ['MD', 'TO', 'RP', 'VB', 'VBN', 'VBD']

  - `PROHIBIT_AS_FINAL`: ['CC', 'DT', 'EX', 'MD', 'PDT', 'POS', 'WDT', 'WP', 'WP$', 'WRB']

  Whenever the segmented sentence regards either case, it is concatenated with the previous sentence or the following sentence. `PROHIBIT_AS_FINAL` was primarily applied.

The list of sentences obtained from the above process is translated by the trained MT module.

5 Experiment

Overall, our speech translation pipeline has the following procedure.

1. Voice segmentation
2. Automatic speech recognition
3. Sentence concatenation
4. Machine translation
5. Checking the performance

Voice segmentation was done separately in the whole pipeline. ASR was performed with 1 RTX 3090. DeepSegment and sentence concatenation were performed with 1 RTX TITAN. MT was performed with 1 RTX 3090. The performance of each trial was checked with the BLEU score.

We achieved the performance of BLEU 20.37 with the official dev set. We finally obtained the performance of BLEU 20.9 with the test set using given segmentation.

6 Conclusion

In this paper, we report the VUS ASR-MT pipeline system for en-de speech translation. The featured engineering schemes are wav2vec-based ASR module, Transformer-based MT, speech segmentation and post-processing, and various cleansing for the enhancement. We obtained similar performance with both dev and test set, the BLEU score of 20.37 and 20.9 respectively. Our model is explainable and partially improvable, given the transparent description of our pipeline system.

Acknowledgments

We thank anonymous reviewers for helpful feedbacks.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

[4] https://github.com/notAI-tech/deepsegment

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*, pages 44–54, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth Heafield, Rohan Kshirsagar, and Santiago Barona. 2015. Language identification and modeling in specialized hardware. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 384–389, Beijing, China. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8417–8424.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Jan Niehues, Roldano Cattoni, Sebastian Stuker, Mauro Cettolo, Marco Turchi, and Marcello Federico. The iwslt 2018 evaluation campaign. In *IWSLT 2018*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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.

Tomasz Potapczyk and Pawel Przybysz. 2020. SR-POL's system for the IWSLT 2020 end-to-end speech translation task. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 89–94, Online. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.