End-to-end Simultaneous Speech Translation with Style Tags using Human Simultaneous Interpretation Data

Yuka Ko[1] Ryo Fukuda[1] Yuta Nishikawa[1] Yasumasa Kano[1]
Katsuhito Sudoh[1,2] Sakriani Sakti[1] Satoshi Nakamura[1,3]
[1]Nara Institute of Science and Technology [2]Nara Women's University
[3]The Chinese University of Hong Kong, Shenzhen
ko.yuka.kp2@is.naist.jp

掲載号の情報

32 巻 2 号 pp. 404-437.
doi: https://doi.org/10.5715/jnlp.32.404

概要

Simultaneous speech translation (SimulST) translates speech incrementally, requiring a monotonic input-output correspondence to reduce latency. This is particularly challenging for distant language pairs, such as English and Japanese, as most SimulST models are trained using offline speech translation (ST) data, where the entire speech input is observed during translation. In simultaneous interpretation (SI), a simultaneous interpreter translates source language speech into target language speech without waiting for the speaker to finish speaking. Therefore, the SimulST model can learn SI-style translations using SI data. However, owing to the limited availability of SI data, fine-tuning an offline ST model using SI data may result in overfitting. To address this problem, we propose an efficient training method for the speech-to-text SimulST model using a combination of small SI and relatively large offline ST data. We trained a single model with mixed data by incorporating style tags to instruct the model to generate either SI or offline-style outputs. This approach, called mixed fine-tuning with style tags, can be extended further using the multistage self-training approach. In this case, we use the trained model to generate pseudo-SI data. Our experimental results for several test sets demonstrated that our models trained using mixed fine-tuning and multistage self-training outperformed baselines across various latency ranges.