ACL Paper Summary

CS 4395 - Noah Gonzales

Learning When to Translate for Streaming Speech

Learning When to Translate for Streaming Speech was written by Qianqian Dong, Yaoming Zhu, Mingxuan Wang, and Lei Li from ByteDance AI Lab at the University of California, Santa Barbara.

The paper focuses on finding the proper moments to generate partial sentence translation for streaming speech. Moments to translate are dynamically decided for real time scenarios, such as for meetings, broadcasts, and interviews. Existing methods, such as waiting then translating for a certain duration, do not translate along acoustic units of speech and often break up acoustic units in speech, making these methods less accurate.

Most prior work involves simultaneous text translation. One work proposes a wait-*k* strategy based on a prefix-to-prefix framework. This method is not very efficient and not accurate for real-time translation due to its fixed segmentation. Other work was done using an end-to-end streaming translation, which also suffered from fixed segmentation.

This paper describes a new system, named the Monotonic-segmented Streaming Speech Translation (MoSST). Specifically, the paper describes a new module that more accurately judges the acoustic boundaries of input audio that works better for streaming audio. This model segments waveforms into acoustic segments. The paper then describes a new translation strategy that enables the new model to determine whether to read audio or write tokens given the audio prefix read in.

MoSST was trained and evaluated on the MuST-C dataset. The MuST-C dataset is a multilingual streaming translation corpus with source audio, transcripts, and text translations. At the time of this paper, it is the largest streaming translation dataset available. The dev and tst-COMMON sets are the development and test data used, respectively. Evaluation was done

in two areas: translation quality and real-time performance. Offline translation was mainly done with quality metrics, and streaming translation was evaluated by the latency-quality trade-off curves. Translation quality was measured using the BLEU metric, which is a standard metric for measuring the similarity between generated translation and reference translations. Latency metrics were evaluated using several metrics: Average Proportion, Average Latency, and Differentiable Average Lagging. The MoSST model was compared to published work for streaming translation and performed better in all three latency metrics and quality metrics. MoSST was also compared to published work on offline streaming translation tasks and achieved an improvement over the best results published so far, obtaining an improvement in quality of 1.3 and 0.7 BLEU with bilingual and multilingual settings respectively.

Qianqian Dong has received 434 citations, Yaoming Zhu has received 648 citations, Mingxuan Wag has received 2118 citations, and Lei Li has received 69531 citations. This paper has received six citations. Although six is not a large amount, the importance of this work is important when considering the importance of live translation. Post-Covid, online forms of communication, such as Zoom, Microsoft Teams, Google Meet, and others have increased in use and have become the standard in some areas for communication. Real-time translation is even more prevalent in business and communication and this work improves translation quality and latency.