Preventing Over-translation in Simultaneous Neural Machine Translation

Name: Yasumasa Kano
Laboratory's name: Augmented Human Communication Laboratory
Supervisor's name: Satoshi Nakamura

Abstract

As globalization expands, a lot of international conferences are held, and we sometimes prefer simultaneous interpretation which starts to translate before the speaker finishes the utterance. Hiring interpreters is highly costly, so there is a huge demand for simultaneous machine translation which can be used at a lower cost. A lot of previous research on simultaneous translation uses the translation model pre-trained with bilingual full-sentence pairs. However, at the inference step, the model needs to translate only a partial input much shorter than a full sentence. This causes the problem of over-translation which outputs a translation longer than the reference. To tackle this problem, we propose to fine-tune the pre-trained translation model with bilingual prefix pairs. A prefix is the initial portion of a sentence. By fine-tuning, we mitigate the gap between training and inference of simultaneous translation and prevent over-translation.

For the evaluation of simultaneous translation models, we need to measure latency in addition to the quality of translation. Latency is the time lag between the input utterance and its translation. When over-translation happens and partial translation becomes longer, it will delay the start of the next translation. However, most existing latency metrics focus on the starting time of translation but do not sufficiently consider the delay caused by the ending time of the previous partial translation. Therefore, we proposed a novel latency metric called Average Token Delay (ATD), which also focuses on the ending time of partial translation.

In our simultaneous machine translation experiments on English-Japanese and English-German, the proposed method improved the quality-latency trade-off of simultaneous translation in low latency for both language pairs. We also compared the correlation of latency metrics and Ear-Voice Span (EVS) which is a latency metric used in human interpretation research, and ATD had the highest correlation with EVS in most conditions in our experiment.