# 論 文 内 容 の 要 旨

博士論文題目

Preventing Over-translation in Simultaneous Neural Machine Translation
（同時ニューラル機械翻訳における過剰翻訳の防止）

氏　　名　　加納 保昌

（論文内容の要旨）

Simultaneous machine translation is an important technology for smooth communication among international people. Although machine translation research has evolved rapidly, simultaneous translation is still a difficult task.

Most 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. Simultaneous translation has a quality-latency trade-off; If the latency is smaller, the model cannot use the latter part of the input sentence and the translation quality would be lower, and vice-versa. When over-translation happens and partial translation becomes longer, it will delay the start of the next translation. However, most of the 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. Average Lagging (AL) is one of the most commonly used latency metrics, but it even gives smaller latency for such long partial translation output. 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, we evaluated the proposed fine-tuning method using AL and ATD. As a result, the proposed method improved the quality-latency trade-off of simultaneous translation in low latency for both language pairs. However, in the high latency ranges in English-Japanese, the proposed method did not work well. We analyzed the reason from the perspective of word order difference.

We also verified the effectiveness of ATD through simulations and analyses us- ing Ear-Voice Span (EVS) which is a latency metric used in human interpretation research. We compared the correlation of latency metrics and EVS, and ATD had the highest correlation with EVS in most conditions in our experiment.

| 氏　名 | 加納 保昌 |
|---|---|

（論文審査結果の要旨）

This thesis tackles the over-translation problem in simultaneous machine translation (SimulMT), where partial inputs are translated into partial outputs to reduce latency in translation. SimulMT models are usually trained using sentence-level machine translation (MT) corpora, so they do not necessarily work for translating partial inputs and often result in over-translation. The over-translation does not only hurt the translation quality but also increases the latency of the translation outputs. The contribution of this thesis is two-fold: the development of a novel data augmentation framework called Prefix Alignment (PA) and the proposal of a novel latency metric for SimulMT called Average Token Delay (ATD). PA extracts pairs of input and output prefixes from bilingual corpora for effective training of SimulMT models. ATD focuses on the latency caused by delays caused by excessively long outputs, which are not taken into account in existing metrics such as Average Lagging. The experimental results revealed the effectiveness of PA and ATD in SimulMT, and detailed analyses were conducted to investigate the advantages of the proposed methods.

This thesis contributed to the progress of SimulMT research from novel perspectives about over-translation in SimulMT. This research resulted in one peer-reviewed journal article and two peer-reviewed international conference papers, and ATD has been used as an official latency metric in the IWSLT shared task. As a result, this thesis sufficiently qualified as a Doctoral thesis of Engineering.