Streaming Neural Machine Translation

Javier Iranzo-Sánchez
AppTek GmbH, València, Spain
jiranzo@apptek.com

Thesis Summary

Speech Translation (ST) is a subfield of Machine Learning (ML) that aims to automatically generate the text translation of a given audio waveform. Currently, the majority of the work in ST is concerned only with the offline task, that is, the task in which the entire input audio is available, and no real-time constraints exist. In contrast, in the online task the input audio is incrementally received as time passes, and the system must produce a translation of a partial input within a certain latency threshold, in a real-time fashion. Online ST is inherently a harder problem, because the partial input compromises the quality of the translation, and due to the need for real-time translation, the computational efficiency of the system cannot be ignored.

Traditionally, ST systems follow the cascade approach, in which the output of an Automatic Speech Recognition (ASR) system is fed into a Machine Translation (MT) system. Direct models are a more recent development, in which a single model receives the audio signal and generates the translation. The techniques presented on this thesis follow the cascade approach, but they can also be applied to the direct approach. Both approaches had achieved a similar level of performance at the time the thesis was written.

This thesis focuses on Streaming ST [1], a subtask of online ST in which the input is an unbounded audio stream. Streaming ST presents additional difficulties when compared with the standard online setup, and it is especially relevant because many potential ST applications such as live lectures or simultaneous interpretation fall under the umbrella of streaming ST. The main goal of this thesis is to develop the tools and techniques that are required in order to create a working streaming ST solution. These are, specifically, a dataset for training and evaluating the ST models, a segmenter system that connects the output of the ASR system with the MT system, a streaming-ready evaluation metric and a streaming-specific MT model that can take advantage of contextual information.

The first challenge is the data scarcity problem faced when training ST systems. In order to alleviate this, a ST dataset is constructed starting from the official recordings of the proceedings of the European Parliament. The data is organized in triples, containing the audio jointly with its transcription and translation. It is a multilingual dataset with 10 different official European languages available both on the source and target side. Document-level information and metadata is included so that this dataset can be used for streaming ST.

The segmentation step is the next challenge to be addressed. The output of the streaming ASR system is a continuous stream of words, which needs to be segmented into semantically self-contained units to be translated by the MT system. We introduce a novel neural segmenter architecture, Direct Segmentation (DS), which considers the segmentation process as a classification problem. Using a sliding window approach, for every position of the ASR stream, the segmenter decides whether or not to produce a chunk by using a fixed local history and a small look-ahead window. The proposed architecture is computationally efficient while outperforming other segmentation approaches, and is able to work straight out

© 2024 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

[1] Streaming ST is also known in the literature as *long-form simultaneous ST*

of the box in the streaming scenario. Experiments are also performed showing that adding audio features to the segmenter improves performance. This work is then extended in order to evaluate the real latency for a simultaneous ST system that uses online ASR and MT systems as well as the proposed DS system. The results show how an acceptable translation quality can be reached at the same latency as a human interpreter (approximately 4 seconds).

The next challenge of streaming ST lies in how to actually evaluate the latency of the ST system under streaming conditions. This thesis introduces a novel evaluation procedure for streaming MT. Standard online MT metrics only work with short audio segments, evaluated in isolation, and do not take into account the sequential nature of the streaming scenario. Our proposed streaming evaluation method fixes these issues, and as a bonus, it can be applied to the standard metrics used for online MT with a small modification. Our proposal keeps track of a global latency score across the entire translation process, and uses a realignment step that matches translated words with the correct reference segment. A significant advantage of our proposal is that the evaluation procedure is not system/segmentation dependent and can be used to compare different systems, as well as maintaining the original interpretability of the metrics. Comparative experiments show that, unlike competing approaches, our proposal correctly ranks systems based on their latency, as well as keeping the previously mentioned properties.

Last but not least, we present a general methodology for building context-aware state-of-the-art streaming MT systems. This approach uses the insights developed in the previous publications in order to build a strong streaming baseline MT system, and improves it with a novel context-aware training methodology which obtains significant improvements. Further improvements are also obtained with a proposed Partial Bidirectional Encoder that has access to a larger portion of the input prefix. Our approach is similar to the concatenative approach used in context-aware MT, and uses a sliding window which contains the previous streaming history that has been produced during the translation process. History-augmented training samples are constructed from document-level corpora, and at inference time, the real streaming history is used. Extensive experiments show how this approach achieves state-of-the-art results.

The full text of the thesis can be accesed at https://doi.org/10.4995/Thesis/10251/199170.

Supervisor Contact Details

Jorge Civera: Associate Professor, Universitat Politècnica de València, València, Spain. jorcisai@vrain.upv.es

Alfons Juan: Full Professor, Universitat Politècnica de València, València, Spain. ajuanci@vrain.upv.es

Acknowledgments

The author would like to publicly acknowledge the support received from his supervisors Jorge Civera and Alfons Juan before, during and after this PhD degree. Likewise, the author received many insightful comments from the thesis committee and external evaluators, consisting of Francisco Casacuberta Nolla, Jesús Andrés Ferrer, Marco Turchi, Felipe Sánchez Martínez and Marta Ruiz Costa-Jussà. This thesis has been developed with the financial support of the FPU scholarship program of the Government of Spain (FPU18/04135), EU's Horizon 2020 project X5gon (761758), as well as Government of Spain's Multisub (RTI2018-094879-B-I00) and Erasmus+ EXPERT (no. 20-226-093604-SCH) research projects. The author gratefully acknowledges the computer resources at Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).