Towards Machine Speech-to-speech Translation

Abstract

There have been lots of research activities on machine speech-to-speech translation (S2ST) in Japan. This article introduces those activities and our recent research activities towards automatic simultaneous speech translation. S2ST system is basically composed of three modules: Large vocabulary continuous automatic speech recognition (ASR), machine text-to-text translation (MT), and text-to-speech synthesis (TTS). All the modules need to be multi-lingual in nature and thus necessitate multi-lingual speech and corpora for training models. The S2ST performance is drastically improved by deep learning and large training corpora. But there still remain many issues such as simultaneity, para-linguistics, context and situation dependency, intention, and cultural dependency. We will introduce the current research activities and discuss issues toward next-generation speech-to-speech translation.

1 Introduction

The drastic increase of demands for cross-lingual conversations, triggered by IT technologies such as the Internet and an expansion of borderless community boosts research activities on a machine speech-to-speech translation (S2ST) technology.

S2ST system is basically composed of three modules: Large vocabulary continuous speech recognition (ASR), machine text-to-text translation (MT), and text-to-speech synthesis (TTS). All the modules need to be multi-lingual for worldwide users and thus necessitate multi-lingual speech and corpora for training models.

On the contrary to the normal machine translation of texts, speech translation receives speech as input and will be sed for online human to human communication. S2ST needs to preserve paralinguistic information such as emotion, emphasis, prominence, and prosody of the source language into speech in the target language. Also, the spoken language needs to consider contexts since people utter not in a complete sentence but in incomplete phrases. Finally, S2ST needs to work in real-time with very low latency and efficiency since it will be used for online real-time communication.

The difficulties of S2ST also depend on similarity of source and target languages. S2ST between western language and non-western languages such as English-from/to-Japanese, English-from/to-Chinese requires different technologies to overcome their drastic difference regarding linguistic expressions. For example, a translation from Japanese to English requires, (1) word separation process for Japanese because Japanese has no explicit spacing information, (2) translating Japanese into English in the drastically different style because their word ordering and their coverage of words are completely different.

2 S2ST Research in Japan

The fist S2ST research project was launched in 1986 in order to overcome the language barrier problem at ATR Interpreting Telephony Research Laboratories in Japan funded by the Ministry of Post and Telecommunication. Afterwards, S2ST research was carried out at ATR until 2008 and at the National Institute of Communication and Technology (NICT), Japan after 2008. Currently developments and deployments of S2ST technologies to the real services for daily conversation such as VoiceTra®[1] are carrying out under the Global Communication Project funded by Ministry of Internal Affairs and Communication.

Research activities on simultaneous speech-to-speech translation at Nara Institute of Science and Technology (NAIST) launched in 2011

[1] https://voicetra.nict.go.jp/en/index.html

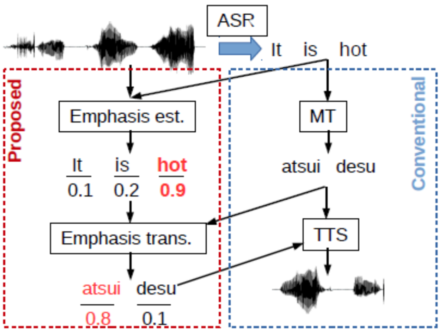

Figure 1: An illustration of the emphasis S2ST. (Do et al., 2018)

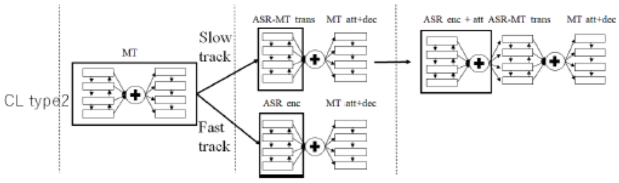

Figure 2: An illustration of the direct S2ST. (Kano et al., 2017)

when the 1st author of this paper moved from NICT to NAIST. We are working on various new challenges for S2ST not just taking ASR outputs as MT inputs, including para-linguistic speech-to-speech translation (PLS2ST), direct speech-to-speech translation, simultaneous speech-to-speech translation (SS2ST) , evaluation of and corpus collection of simultaneous interpretation are going on at NAIST. This paper introduces those activities in the following sections.

3 Our research activities

3.1 Para-linguistic speech translation

For the transfer of para-linguistic information of emphasis, we have proposed a method based on encoder-decoder with attention (Do et al., 2018). This method estimates emphasis in the source speech and maps it into the target speech within the encoder-decoder cascaded speech-to-speech translation framework. Figure 1 illustrates the para-linguistic speech translation system. This framework will be extended to incorporate emotions in the future.

3.2 Direct speech translation

Another attempt is to realize direct speech-to-speech translation to translate linguistic and para-linguistic information into one framework. We have proposed a method using curriculum learning based on encoder-decoder direct speech translation (Kano et al., 2017). The neural network architectures have been shown to provide a powerful model for machine translation and speech recognition. Recently, several studies have attempted to extend the models for end-to-end speech translation tasks. However, the usefulness of these models was only investigated on language pairs with similar syntax and word order (e.g., English-French or English-Spanish). We proposed an end-to-end speech translation model for syntactically distant language pairs (e.g., English-Japanese) that require distant word reordering (Kano et al., 2017). To guide the encoder-decoder attentional model to learn this difficult problem, we propose a structured-based curriculum learning strategy starting from independently-trained modules and then fine-tuning the overall network. Also, we introduced a neural transcoder to convert ASR decoder outputs to MT encoder outputs. We start the training with end-to-end encoder-decoder for speech recognition or text-based machine translation tasks then gradually move to end-to-end speech translation task. The experiment results confirmed that our proposed approach could provide significant improvements.

3.3 Simultaneous speech translation

Simultaneous interpretation is a very challenging task in human verbal communication that requires strong expertise. We are trying to mimic this simultaneous process through computers using speech translation technologies. We call it *simultaneous speech translation* since the current machine translation is really far from *interpreting* human languages. The most significant challenge here is the latency between the input speech and translated output especially in syntactically distant languages such as English (Subject-Verb-Object) and Japanese (Subject-Object-Verb).

3.3.1 Latency in simultaneous translation[2]

Suppose we are going to translate the following English sentence into Japanese (Mizuno, 2016).

[2] Materials in this section are from Mizuno (2016).

> (1) The relief workers (2) say (3) they don't have (4) enough food, water, shelter, and medical supplied (5) to deal with (6) the gigantic wave of refugees (7) who are ransacking the countryside (8) in search of the basics (9) to stay alive.

In a standard English-to-Japanese translation, we translate the sentence almost in a reverse order based on syntactic correspondence in Japanese.

> Example 1: (1) *Kyūen tantōsha tachi ha* [The relief workers] (9) *ikirutame no* [to survive] (8) *shokuryō wo motomete* [in search of food] (7) *mura wo arashi mawatte iru* [are ransacking the countryside] (6) *tairyō no nanmin tachi no* [a healthy number of refugees] (5) *sewa wo suru tameno* [to take care of] (4) *jūbun na shokuryō ya mizu, shukuhaku shisetsu, iyakuhin ga* [sufficient amount of food, water, lodgings, and medical supplies] (3) *nai to* [don't have] (2) *itte imasu* [are saying].

The chunk-level correspondence and memory load are shown in Figure 3. The chunks (2) to (9) are stored in the memory to translate them with the correct syntactic structure in Japanese. As a result, the ear voice span becomes very large, and that makes the interpretation process difficult; next inputs will come even when an interpreter speaks. Furthermore, it is tough to maintain so many numbers of chunks for interpreters.

On the other hand, Mizuno (2016) presented an interpretation example with an ideal strategy with monotonic translation as follows.

> Example 2: (1) *Kyūen tantōsha tachi no* [The relief workers] (2) *hanashi de ha* [according to their talk] (4) *shokuryō, mizu, shukuhaku shisetu, iyakuhin ga* [food, water, shelters, and medical supplies] (3) *tarizu* [are in short supply] (6) *tairyō no nanmin tachi no* [a massive amount of refugees] (5) *sewaga dekinai tono kotodesu* [cannot be taken care of]. (7) *Nanmin tachi ha ima muramura wo arashi mawatte* [The refugees are now ransacking the villages] (9) *ikiru tameno* [to survive] (8) *shokuryō wo motomete irunodesu* [searching for the basics].

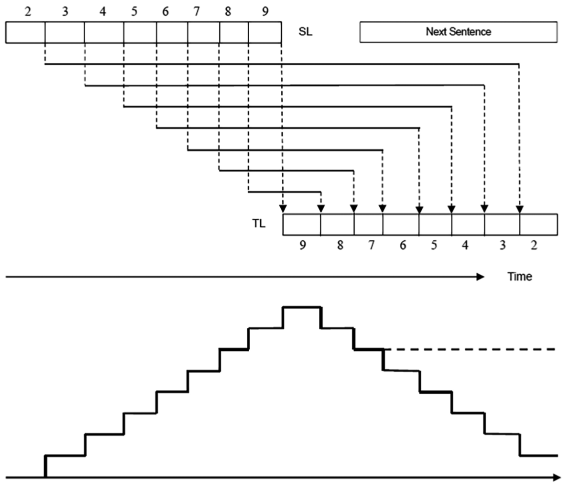

Figure 3: Chunk correspondence in Example 1 (chunk 1 should be translated at first and ignored in this diagram). SL and TL stands for source language and target language, respectively. The chart below illustrates the corresponding chunk-level memory load.

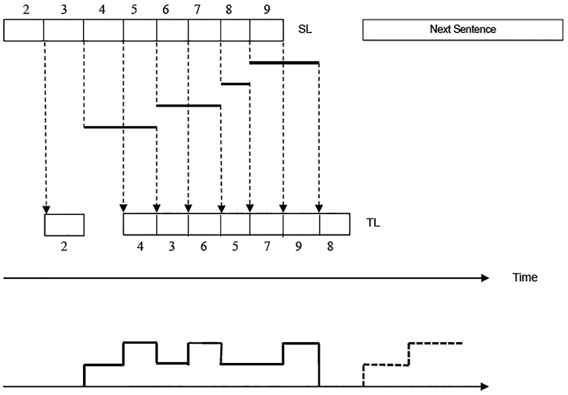

Figure 4: Chunk correspondence in Example 2.

This example has two substantial differences from Example 1. First, the main verb, *say*, is translated immediately, and the following contents are translated after it. Second, the relative clause starting from *who* is translated after its modifiee chunk (6) as follow-up information. Using this kind of *monotonic* interpretation, the latency and memory load become much smaller than the previous example, as shown in Figure 4. One of the most important challenge for simultaneous S2ST is this kind of monotonic translation, as experienced interpreters do.

3.3.2 Automatic simultaneous S2ST

We are working on time-synchronous and incremental processing in ASR, MT, and TTS for small latency S2ST, using recent neural network (NN) technologies. We proposed a NN-based incremental ASR method (Novitasari et al., 2019). The proposed method focuses only on very recent parts of speech inputs, while a standard NN-based ASR looks over an utterance. In our experiments, the proposed method reduced the transcription errors with allowing a delay in 400 msec. to include some context information into ASR. With respect to the MT, we proposed an incremental neural MT method (Chousa et al., 2019). In simultaneous S2ST, this MT part has the largest effect on overall latency, because we can easily face a seriously long delay as discussed above. The proposed method can wait for future inputs when we are not confident of translation based on currently observed inputs. Finally, for the TTS, we proposed an incremental neural TTS method (Yanagita et al., 2019). The TTS model of the proposed method is trained using short segments of text-speech pairs, and we use the model to synthesize speech signals at the segment level. In our experiments, allowing a delay in two to three words contributed the synthesized speech quality.

3.3.3 Corpus development

We are developing a simultaneous translation corpus for our simultaneous S2ST research. The corpus includes recordings of simultaneous interpretations by professional interpreters with different experiences (S: more than 15 years, A: 4 years or more, B: less than 4 years). Currently, we have about 150 hours of English-to-Japanese and 100 hours of Japanese-to-English interpretations with transcriptions, mostly in lecture talks like TED Talks. Such a large scale simultaneous interpretation corpus in Japanese-English does not exist so far. We are going to accelerate our research on simultaneous S2ST with this corpus.

4 Summary

This article summarized our research activities on the S2ST system. The S2ST performance is drastically improved by deep learning and large training corpora and the deployment to the real services like VoiceTra has been started. But there still remain many issues such as simultaneity, paralinguistics, context and situation dependency, intention, and cultural dependency. Further fundamental research is necessary to overcome those problems toward natural speech-to-speech translation which resembles more closely the output of human interpreters.

References

Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. 2019. Simultaneous Neural Machine Translation using Connectionist Temporal Classification. *arXiv preprint*, 1911.11933.

Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. 2018. Sequence-to-Sequence Models for Emphasis Speech Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1873–1883.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2017. Structured-Based Curriculum Learning for End-to-End English-Japanese Speech Translation. In *Proceedings of Interspeech 2017*, pages 2630–2634.

Akira Mizuno. 2016. Simultaneous Interpreting and Cognitive Constraints. *Journal of College of Literature, Aoyama Gakuin University*, 58:1–28.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition. In *Proceedings of Interspeech 2019*, pages 3835–3839.

Tomoya Yanagita, Sakriani Sakti, and Satoshi Nakamura. 2019. Neural iTTS: Toward Synthesizing Speech in Real-time with End-to-end Neural Text-to-Speech Framework. In *Proceedings of the 10th ISCA Speech Synthesis Workshop*, pages 183–188.