Towards Machine Speech-to-speech Translation

Nakamura Satoshi
Katsuhito Sudoh
Sakriani Sakti

revista.tradumàtica
tecnologies de la traducció

Nakamura Satoshi
Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan
s-nakamura@is.naist.jp;
ORCID:
0000-0001-6956-3803

Katsuhito Sudoh
Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan
sudoh@is.naist.jp;
ORCID:
0000-0002-2122-9846

Sakriani Sakti
Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan
ssakti@is.naist.jp;

Abstract

There has been a good deal of research on machine speech-to-speech translation (S2ST) in Japan, and this article presents these and our own recent research on automatic simultaneous speech translation. The 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 these modules need to be multilingual in nature and thus require multilingual speech and corpora for training models. S2ST performance is drastically improved by deep learning and large training corpora, but many issues still still remain such as simultaneity, paralinguistics, context and situation dependency, intention and cultural dependency. This article presents current on-going research and discusses issues with a view to next-generation speech-to-speech translation.

Keywords: Speech-to-speech translation, automatic speech recognition, machine text-to-text translation, text-to-speech synthesis.

Resum

Al Japó s'han dut a terme moltes activitats de recerca sobre la traducció automàtica de la parla. Aquest article n'ofereix una visió general i presenta les activitats que s'han efectuat més recentment. El sistema S2ST es compon bàsicament de tres mòduls: el reconeixement automàtic de la parla contínua i de vocabularis extensos (Automatic Speech Recognition, ASR), la traducció automàtica de textos (Machine translation, MT) i la conversió de text a veu (Text-to-Speech Synthesis, TTS). Tots els mòduls han de ser plurilingües, per la qual cosa es requereixen discursos i corpus multilingües per als models de formació. El rendiment del sistema S2ST millora considerablement per mitjà d'un aprenentatge profund i de grans corpus formatius. Tanmateix, encara cal tractar diversos aspectes, com la simultaneïtat, la paralingüística, la dependència del context i de la situació, la intenció i la dependència cultural. Així, farem un repàs a les activitats de recerca actuals i discutirem diverses qüestions relacionades amb la traducció automàtica de la parla d'última generació.

Paraules clau: Traducció automàtica de la parla, reconeixement automàtic de la parla, traducció automàtica de textos, conversió de text a veu.

Rebuda: 29 de novembre de 2019 / Acceptació: 16 de desembre de 2019 / Publicació avançada: 23 de desembre de 2019

Revista Tradumàtica. Número 17 (desembre 2019).
Cita recomanada: SATOSHI, NAKAMURA; SUDOH, KATSUHITO; SAKTI, SAKRIANI (2019). «Towards Machine Speech-to-speech Translation». Revista Tradumàtica. Tecnologies de la Traducció, 17, 81-87. <https://doi.org/10.5565/rev/tradumatica.238>

http://revistes.uab.cat/tradumatica
ISSN 1578-7559

Resumen

En Japón se han llevado a cabo muchas actividades de investigación acerca de la traducción automática del habla. Este artículo pretende ofrecer una visión general de dichas actividades y presentar las que se han realizado más recientemente. El sistema S2ST está formado básicamente por tres módulos: el reconocimiento automático del habla continua y de amplios vocabularios (Automatic Speech Recognition, ASR), la traducción automática de textos (Machine translation, MT) y la conversión de texto a voz (Text-to-Speech Synthesis, TTS). Todos los módulos deben ser plurilingües, por lo cual se requieren discursos y corpus multilingües para los modelos de formación. El rendimiento del sistema S2ST mejora considerablemente por medio de un aprendizaje profundo y grandes corpus formativos. Sin embargo, todavía hace falta tratar diversos aspectos, com la simultaneidad, la paralingüística, la dependencia del contexto y de la situación, la intención y la dependencia cultural. Por todo ello, repasaremos las actividades de investigación actuales y discutiremos varias cuestiones relacionadas con la traducción automática del habla de última generación.

Palabras clave: Traducción automática del habla, reconocimiento automático del habla, traducción automática de textos, conversión de texto a voz.

1. Introduction

The major increase in demand for cross-lingual conversations, triggered by IT technologies such as the Internet and an expanding borderless community, has fuelled research into machine speech-to-speech translation (S2ST) technology. The 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 these modules need to be multilingual for users around the world and thus necessitate multilingual speech and corpora for training models.

As opposed to the machine translation of texts, speech translation receives verbal input to be expressed orally in online human-to-human communication. Firstly, S2ST needs to preserve the source language paralinguistic information, such as emotion, emphasis, prominence and prosody, in the target language speech. Secondly, the spoken language needs to consider contexts since utterances do not tend to be in compete sentences but rather incomplete phrases. Finally, S2ST needs to work in real-time with very low latency and efficiency since it will be used for real-time communication online.

From another perspective, S2ST difficulties also depend on the degree of similarity between source and target languages. S2ST between western and non-western languages such as English-from/to-Japanese, English-from/to-Chinese, requires different technologies to overcome the major linguistic differences. For example, translating from Japanese to English requires, (1) word separation for Japanese because Japanese has no explicit spacing information, (2) translating Japanese into English involves a completely different style due to word order and their coverage.

2. S2ST Research in Japan

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

Research activities on simultaneous speech-to-speech translation at the Nara Institute of Science and Technology (NAIST) was launched in 2011 when Shiroishi Nakamura moved from NICT to NAIST. We are working on various new challenges for S2ST, not just taking ASR outputs as MT inputs, which include: paraglinguistic speech-to-speech translation (PLS2ST), direct speech-to-speech translation, simultaneous speech-to-speech translation (SS2ST) and evaluating a corpus collection of simultaneous interpretation. The sections which follow present these on-going research iniatives.

3. Our research activities

3.1. Paralinguistic speech translation

For the transfer of paraglinguistic information of emphasis, we have proposed a method based on encoder-decoder. 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 paraglinguistic speech translation system. This framework will be enhanced to incorporate emotions in the future.

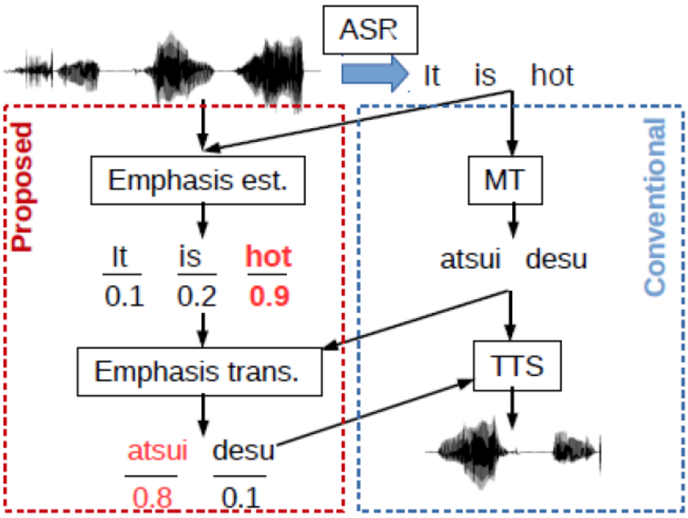

Figure 1: An illustration of the emphasis S2ST

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

3.2. Direct speech translation

Another initiative is to incorporate direct speech-to-speech translation to translate linguistic and paraglinguistic information into one framework. We have proposed a method using curriculum learning based on encoder-decoder direct speech translation. Neural network architectures have been shown to provide a powerful model for machine translation and speech recognition, and several recent studies have attempted to extend the models for end-to-end speech translation tasks. However, the usefulness of these models was only studied for language pairs with similar syntax and word order (e.g., English-French or English-Spanish). We propose an end-to-end speech translation model for syntactically distant language pairs (e.g., English-Japanese) that require distant word reordering. 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.

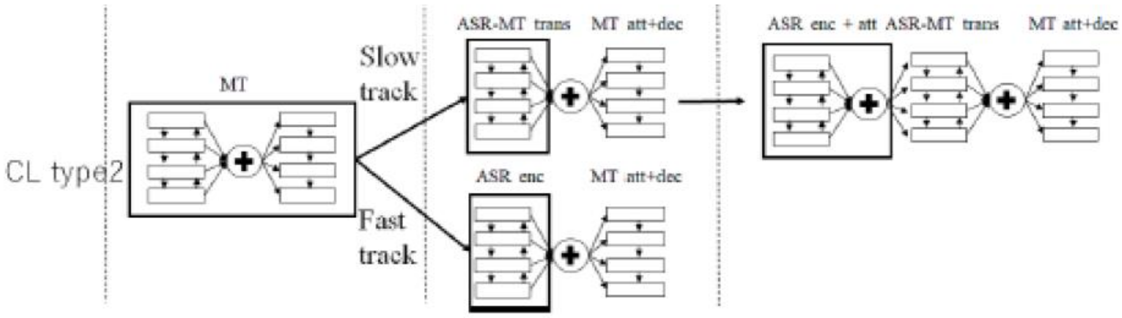

Figure 2: An illustration of the direct S2ST

3.3. Simultaneous speech translation

Simultaneous interpretation is a very challenging task in human verbal communication that requires a high level of expertise. We are trying to mimic this simultaneous process through computers using speech translation technologies. We call this “simultaneous speech translation” since current machine translation is a long way from “interpreting” human languages. The most significant challenge here is the latency between 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.

[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 supplies (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 tanōsha 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 difficult for interpreters to maintain so many numbers of chunks.

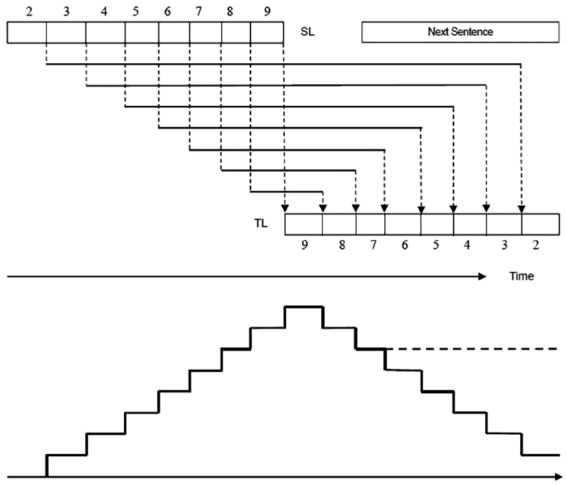

Figure 3: Chunk correspondence in Example 1 (chunk 1 should be translated first and ignored in this diagram). SL (source language) and TL (target language). The chart below illustrates the corresponding chunk-level memory load.

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 shisetsu, 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].

Compared to Example 1, there are two substantial differences here. 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 modified chunk (6) as follow-up information. Using this kind of *monotonic* interpretation, the latency and memory load become much smaller than in the previous example, as shown in Figure 4. One of the most important challenges for simultaneous S2ST is this kind of monotonic translation, as experienced interpreters do.

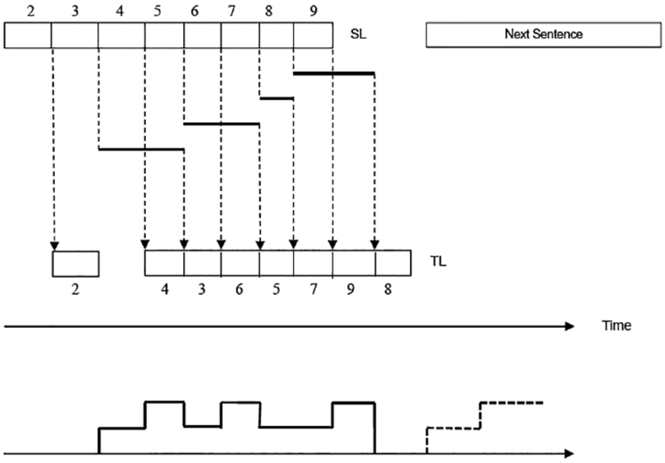

Figure 4: Chunk correspondence in Example 2.

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 propose a NN-based incremental ASR method, which 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 allowing a 400-msec. delay to include some context information into ASR. With respect to the MT, we proposed an incremental neural MT method. In simultaneous S2ST, this MT part has the largest effect on overall latency, because we can easily face a seriously long delay as mentioned above. The proposed method can wait for future inputs when we are not confident of a translation based on currently observed inputs. For the TTS, we propose an incremental neural TTS method; 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 has not been produced so far. The plan is to to accelerate our research on simultaneous S2ST with this corpus.

4. Concluding remarks

This article summarises our research on the S2ST system. S2ST performance is drastically improved by deep learning and large training corpora, and deployment to 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 more closely resemble the output of human interpreters.

References

Chousa, K.; Sudoh, K.; Nakamura, S. (2019). Simultaneous Neural Machine Translation using Connectionist Temporal Classification. ArXiv Preprint, 1911.11933. Retrieved from http://arxiv.org/abs/1911.11933

Do, Q. T.; Sakti, S.; Nakamura, S. (2018). Sequence-to-Sequence Models for Emphasis Speech Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* v. 26, n. 10, pp. 1873–1883. https://doi.org/10.1109/TASLP.2018.2846402

Kano, T.; Sakti, S.; Nakamura, S. (2017). Structured-Based Curriculum Learning for End-to-End English-Japanese Speech Translation, in: *Proceedings of Interspeech 2017*, pp. 2630–2634. https://doi.org/10.21437/Interspeech.2017-944

Mizuno, A. (2016). Simultaneous Interpreting and Cognitive Constraints. *Journal of College of Literature,* Aoyama Gakuin University, n. 58, 1–28. https://www.agulin.aoyama.ac.jp/repo/repository/1000/19723/

Novitasari, S.; Tjandra, A.; Sakti, S.; Nakamura, S. (2019). Sequence-to-Sequence Learning via Attention Transfer for Incremental Speech Recognition, in: *Proceedings of Interspeech* 2019, pp. 3835–3839. https://doi.org/10.21437/Interspeech.2019-2985

Yanagita, T.; Sakti, S.; Nakamura, S. (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, pp. 183–188. https://doi.org/10.21437/SSW.2019-33