Human-like Conversational Robot

Impact of End-to-End Speech Recognition

In the last decade, we have observed a significant progress in speech and image recognition. It was surprising but predictable that deep learning and big data brought this improvement. But more surprising is end-to-end or seq2seq neural network model is replacing the statistical framework of speech recognition, which deemed very solid.

Conventionally, automatic speech recognition (ASR) has been a complex of acoustic model, phone model, lexical model and language model, each of which is formulated with a dedicated statistical model. Currently, a complex but single neural network is designed to conduct the entire process in an end-to-end framework. As illustrated in Figure 1, the direct mapping from a sequence of acoustic features into a sequence of words, called acoustic-to-word model, achieves comparable performance to the state-of-the-art system with an extremely simple architecture and an amazing decoding speed (RTF of 0.04).

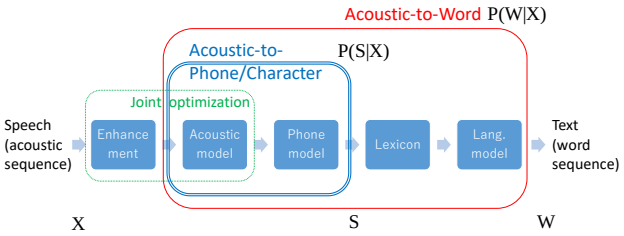

Figure 1 End-to-end speech recognition

End-to-End Dialogue System?

The next step will be end-to-end speech understanding, which directly converts speech into meaning or intentions of the utterance by combining the language understanding function. The question here is how to define a set of meaning and intentions, which may depend on the task domain.

Since the goal of intelligent machines is to respond to any user queries, an ultimate end-to-end model is to output proper responses given a user input. When we assume text for inputs and outputs, the model is called a neural conversational model, which has been intensively investigated in these years. However, we must be aware that dialogue is not a simple mapping from a user input to a system output, as shown in Figure 2. First, we need to take into account the context of the dialogue. The response to "How about you?" cannot be made without knowing the context. Second, we need an external database or knowledge base to respond to many queries. This is not limited to the conventional tasks of database query or question-answering, but even in chatting, we need personal profiles and common senses to make a consistent and relevant dialogue. What is needed is an open problem. Moreover, there are other factors that affects the responses in human dialogue. They include emotions, desire and characters. Moreover, the emotions and desire can change according to the input during the dialogue. These are not modeled in conventional dialogue systems, but would be required in human-like agents and robots.

Another serious problem in training dialog systems is there are many choices in responses given a user input and there is no ground-truth in chatting-style conversations. Modeling emotions, desire and characters will provide a solution. However, we need to define their objective measurement and the evaluation criteria.

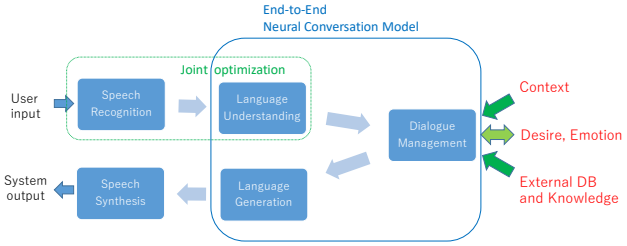

Figure 2 End-to-end spoken dialogue system

Tatsuya Kawahara

PhD, FIEEE

EiC APSIPA T-SIP (2018-), VP-Publications (2014-17)

Professor

School of Informatics

Kyoto University, Japan

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor in the School of Informatics, Kyoto University. He has also been an Invited Researcher at ATR and NICT.

He has published more than 300 technical papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius *and the automatic transcription system for the Japanese Parliament (Diet).*

He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012.

He has been an editorial board member of Elsevier Journal of Computer Speech and Language *and* IEEE/ACM Transactions on Audio, Speech, and Language Processing. *He is a board member of APSIPA and ISCA, and a Fellow of IEEE.*

Android ERICA

Since 2014, we are conducting a project to develop an autonomous android ERICA. Our ultimate goal is to pass a **Total Turing Test**, convincing people that ERICA is indistinguishable from a human in terms of verbal and non-verbal communications including facial expressions and eye-gaze as well as spoken dialogue. This is apparently very challenging even in a 20-year span, as it is almost equivalent to make a human, but we hope this challenge would reveal what is missing in the current technology and what is critical in human communication.

Our realistic goal is to make the interaction with ERICA is as engaging as that with a human. Toward this goal, we set up several social tasks designated for ERICA. Unlike conventional conversational agents and robots, we focus on long and deep interactions, which are not a sequence of short query-response pairs. We choose tasks in which ERICA plays a relatively simple role, focusing on some aspects of interactions.

The first task is **attentive listening**, in which ERICA listens to senior people talking about a given topic such as memorable travels and recent activities. It is similar to counseling, in that she needs to encourage users to talk more by showing interest and empathy. We have investigated generation of natural backchannels in terms of timing, lexical tokens and prosody. We also incorporate partial repeats and elaborating questions based on focus word detection, which can be robustly applied to open-domain conversations.

The second task is **job interview**, in which ERICA asks questions to students applying for a job position in some company. In this scenario, dialogue is designed to be adaptive by generating questions on the fly, without assuming a particular job or a company.

You can watch some demonstration videos of dialogue with ERICA at:
https://www.youtube.com/channel/UCDjRgo5ecEw0Ou78-uJOssg (most of them are in Japanese).
We hope that ERICA can take a role of counselor and interviewer in the future, but the dialogue structure in these roles is relatively simple.

The next task we set up is **speed dating**, which involves asking and answering questions as well as listening and talking. This task calls for the dialogue modeling mentioned in the previous section. We design an emotion model which is affected by the way of talking and listening of the user. It is pre-trained with a set of questionnaires conducted in dialogue data collection, and fine-tuned with dialogue behaviors in an end-to-end manner. We are not sure speed dating with ERICA is pleasant or stressful, but would like to make it real.

Communication skill is one of the fundamental skills of human being and still very important in the current society. For example, face-to-face interview is essential in recruiting students and employees. ERICA is designed not only to replace some human roles, but also to help human practice the skill. This is the reason why we choose job interview and speed dating as target tasks.

Figure 3 Dialogue with ERICA

From Deep Learning to Wide Learning

One of the key concepts of deep learning is an integrated architecture and joint optimization of signal and information processing. This framework has been amazingly successful once we define input and output, and prepare a large amount of the paired data.

While this is regarded as a vertical connection of signal and information processing, we should also explore a horizontal connection that selects relevant information among many sources of signals. This capability is inherent in human being and needed for future AI.

References

[1] T.Kawahara. Spoken dialogue system for a human-like conversational robot ERICA. In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), (keynote speech), 2018.

[2] T.Kawahara, T.Yamaguchi, K.Inoue, K.Takanashi, and N.Ward. Prediction and generation of backchannel form for attentive listening systems. In Proc. INTERSPEECH, pp.2890--2894, 2016.

[3] S.Ueno, H.Inaguma, M.Mimura, and T.Kawahara. Acoustic-to-word attention-based model complemented with character-level CTC-based model. In Proc. IEEE-ICASSP, pp.5804--5808, 2018.

[4] ERICA channel: https://www.youtube.com/channel/UCDjRgo5ecEw0Ou78-uJOssg