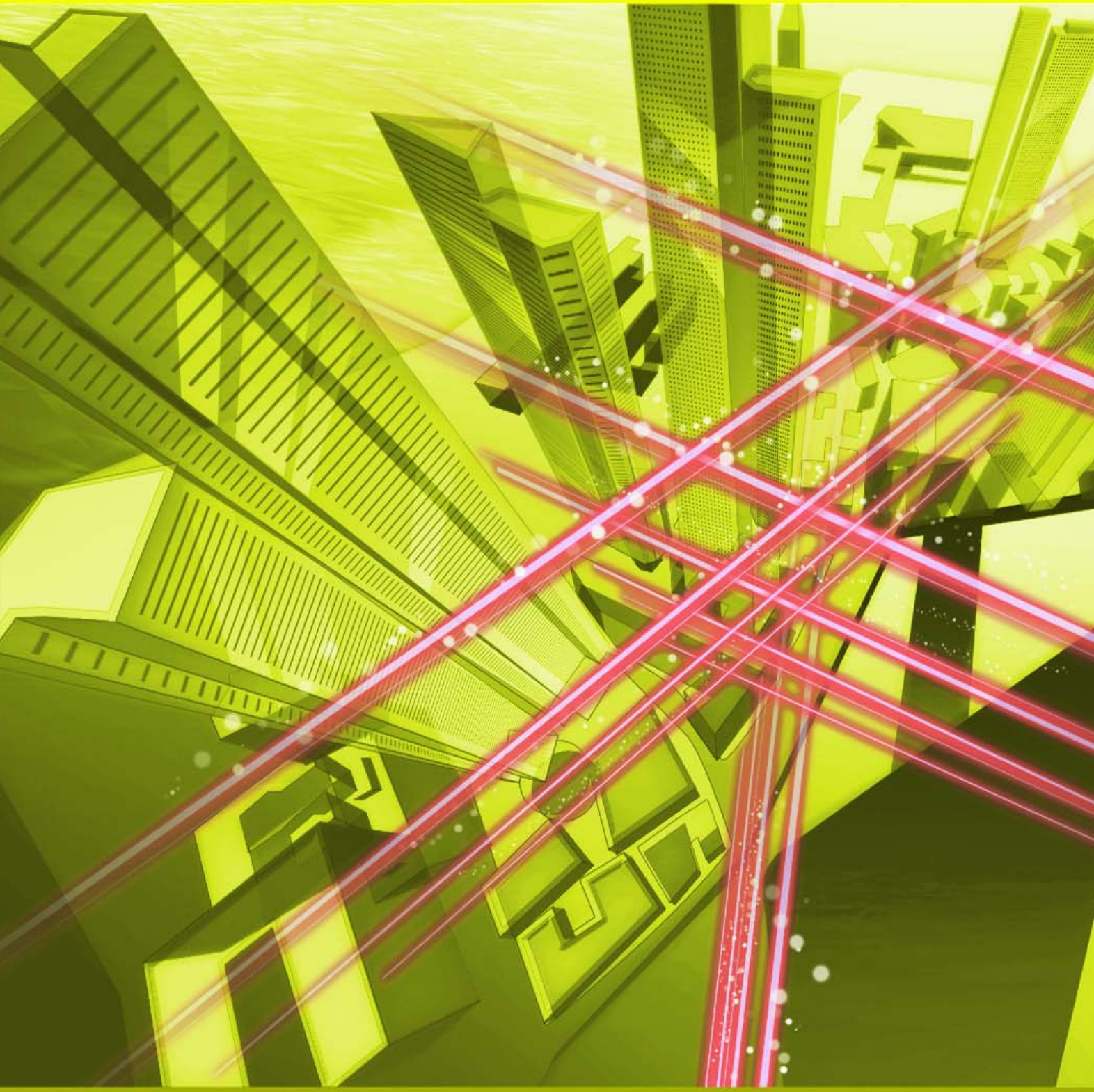
NTT Technical Review

9
2019



September 2019 Vol. 17 No. 9

NTT Technical Review

September 2019 Vol. 17 No. 9

Advanced Initiatives for Contact Center AI

Kimihito Tanaka, Takashi Yagi, and Tetsuya Iizuka

Abstract

Contact centers are becoming increasingly important as a point of contact where feedback from many customers can be obtained. NTT Media Intelligence Laboratories is carrying out research and development of the application of artificial intelligence technology in contact centers. This article introduces some of the latest technologies for solving various issues at contact centers using the speech and natural language processing technologies that we have cultivated over many years.

Keywords: artificial intelligence, contact center, digital transformation

1. Introduction

It has been more than three years since the third artificial intelligence (AI) boom began to take hold in industry, and practical use of AI has begun in various fields. The NTT Group announced its corevo® AI brand in June 2016 and has been advancing various initiatives [1] since then. There is no doubt that the main technology accompanying the third AI boom has been machine learning, and deep learning in particular. Deep learning can be applied as a basic technology in various areas, but in most cases, simple application of the technology will not achieve the best performance when used in actual business. To achieve performance adequate for practical use, a network model specialized for the application must be created and trained using suitable data.

In the early days of AI, there was discussion regarding general AI as opposed to narrow AI. General AI refers to artificial intelligence that is able to solve any type of problem, similar to the human brain. We are still far from achieving this. In contrast, narrow AI can solve particular types of problems, which can be said of all AI available today. By specializing on a particular type of problem, narrow AI is often able to achieve performance equal to or better than humans. As such, an application area must be defined and the AI must be applied to the specific problem to use AI as currently available.

Contact centers are one area where NTT Media Intelligence Laboratories is focusing research and development (R&D) to apply AI technology. Contact centers respond to many telephone calls and chats daily and also conduct many knowledge searches and a lot of call analysis. Thus, it is a field where the speech and natural language processing technologies that we have cultivated for many years can be utilized. There are many contact centers within the NTT Group, and they can provide a lot of data across many fields that can be used in refining our technologies. Our objective is to use these environments and this store of data to create technologies that can be applied in real business scenarios.

2. Contact center issues

Contact center departments specialize in dealing with customers through various channels such as telephone, email, and chat. They were originally positioned as administrative centers for accepting applications or providing customer support, but with recent changes in business environments, they are becoming increasingly important as customer contact points that gather feedback from many customers. The contact center environment in Japan is beset with an increasingly serious shortage of personnel, and hiring and retaining operators is a serious challenge [2]. At the same time, service quality, or customer

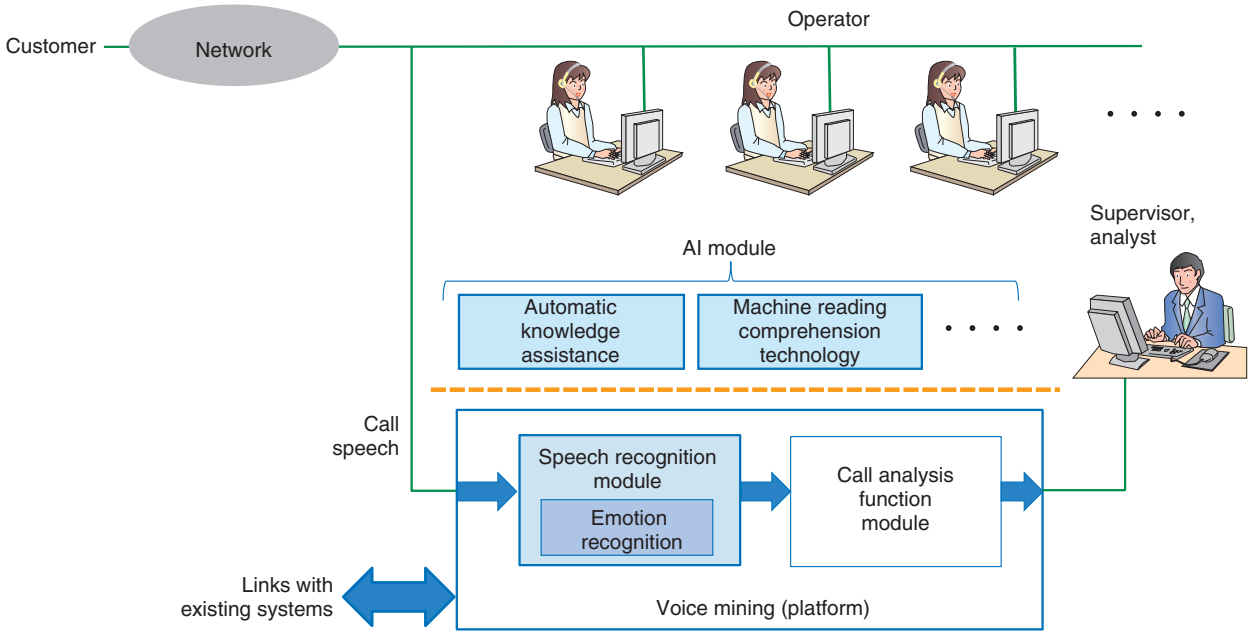


Fig. 1. Example deployment of voice mining technology.

experience (CX), must be improved. Finding ways to improve CX is another major issue for contact center management, especially in centers with high staff turnover and limited operational resources, where often there are personnel shortages and a number of inexperienced staff.

3. AI in contact centers

Various information technology systems have been introduced into contact centers, for example, those for interactive voice response, computer telephony integration, knowledge systems, and customer relationship management. AI is being used to make these systems more sophisticated, to provide new functionality, and to improve operational efficiency and CX.

A typical AI technology used in contact centers is voice mining. Voice mining has achieved dramatic increases in speech recognition accuracy through the use of deep learning, and it is now one of the most widely used AI technologies in contact centers. Through our subsidiary, NTT TechnoCross, NTT is developing and marketing a commercial voice mining technology called ForeSight Voice Mining [3]. It converts call speech to text using speech recognition and applies statistical analysis and visualization from a large volume of calls to obtain hints regarding issues with contact center operation, and how to solve them [4].

Statistical analysis can be used, for example, to extract know-how from skilled operators and deploy it horizontally within the contact center, thereby increasing the overall skill level, or to compile statistical information for the entire center and quickly identify anomalies (such as rapid increases in a particular type of inquiry). Since the speech has been converted to text, it is also easier to review details after a call has been completed, increasing speed and accuracy when entering the call response history. These features are already being used in contact centers within and outside the NTT Group and are achieving results such as improved efficiency and increased sales.

An example of a voice mining technology configuration is shown in **Fig. 1**. The voice mining technology performs speech recognition on the call speech data, and the results are then analyzed by a call analyst. Voice mining has a role as an AI platform in the contact center, and other AI modules can be added to it to use speech recognition results in more sophisticated ways.

The NTT laboratories are also conducting R&D to extract information other than text, for example, emotional content, using AI. Such AI can be used to extend the speech recognition component and expand the scope of analysis. We are also advancing R&D on

AI that creates and presents knowledge based on the results of speech recognition in order to support operators in dealing with inquiries. These types of AI are configured so that they can be built using the voice mining technology as a platform. This enables them to be deployed quickly in real contact center environments.

4. Initiatives for further advancement

Several new issues that need to be addressed have become apparent through commercial deployment of voice mining technology.

(1) Further improvement of customer speech recognition accuracy
(2) Creation of still more value using speech recognition results
(3) Reduction of the cost and time involved in tuning the AI upon introduction

Currently, the accuracy of operator speech recognition in contact centers exceeds 90%. However, accuracy of customer speech recognition is generally ten percentage points lower. The customer speech can include more background noise, depending on their location, and if they are using a mobile phone, the sound quality can be degraded due to encoding and other factors. The speech may also be less stable and less formal than that of the operator, so there are several conditions that can make speech recognition more difficult.

The level of customer speech recognition accuracy is currently adequate for a person to understand what was said from the results and to use the results for statistical analysis at the individual word level. However, higher accuracy is required in order to use results in more sophisticated ways such as for knowledge creation or search. The article, "Evolution of Speech Recognition System—VoiceRex" [5] in the Feature Articles in this issue, introduces some of the latest technologies for resolving issue (1) above.

There is increasing demand when introducing voice mining technology to use speech recognition results in more sophisticated ways, to provide more advanced support for operators, and to reduce costs. Regarding issue (2), the NTT laboratories are focusing their efforts on R&D on knowledge support, which will facilitate acquisition of the knowledge needed for operators to handle inquiries.

The article "Toward Natural Language Understanding by Machine Reading Comprehension" [6] introduces technology that derives answers to questions from manuals and other documents. Knowledge used in contact centers includes materials such as frequently asked questions (FAQs), manuals, and user agreements, but there are many contact centers that have not adequately organized their FAQs.

Machine reading comprehension promises to reduce the cost of organizing such knowledge by eliminating the need to prepare question and answer pairs ahead of time, as is done for FAQ search. In interviews, operators have also expressed opinions such as, "The manuals used in training are usually more familiar than the FAQs and are easier to use," and, "When we have an inquiry that is not in the FAQs, we need to look at the manuals anyway," so there is also increasing anticipation among operators for the convenience such technology will provide.

Another article in this issue, "Automatic Knowledge Assistance System Supporting Operator Responses" [7], introduces an automated knowledge assistance technology that links speech recognition with FAQ search, automatically presenting relevant FAQ entries at the appropriate time to operators while they are handling inquiries. Automatic knowledge assistance eliminates the time required for the operator to search for knowledge. The article also introduces efforts to increase the efficiency of organizing FAQ information using the speech recognition results.

The accuracy of AI technologies such as speech recognition can be increased by tuning them for each application area, so it is important to reduce the cost and time required for tuning, especially when deploying it in small-scale contact centers. For issue (3), the NTT laboratories are working on developing new technologies that improve accuracy and also working on reducing the tuning required for each installation by strengthening the underlying AI models. We have been collecting large amounts of training data from voice calls and other sources and creating industry-specific base models for speech recognition and knowledge search. In the major industries for which we have built industry-specific base models, we are now able to implement highly accurate systems, quickly and at low cost, by simply performing additional training with a small amount of training data.

5. Expanding scenarios for using AI

The accuracy of speech recognition and natural language processing has increased dramatically due to the emergence of deep learning, increases in computer performance, and training with large amounts of data. These technologies are becoming common in

real business environments. In the future, we intend to pursue successful business projects with various partners, using the media processing technologies introduced here and others. We will also continue promoting R&D on innovative technologies to expand the range of applications for AI.

References

[1] Corevo, http://www.ntt.co.jp/corevo/e/index.html
[2] Monthly Call Center Japan Editorial Dept., "Call Center White Paper 2018," RIC Telecom, 2018 (in Japanese).
[3] ForeSight Voice Mining, https://www.ntt-tx.co.jp/eng/products/foresight_vm/
[4] S. Kawamura, K. Machida, K. Matsui, D. Sakamoto, and M. Ishii, "Utilization of Artificial Intelligence in Call Centers," NTT Technical Review, Vol. 14, No. 5, 2016.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201605fa7.html
[5] T. Oba, T. Tanaka, and R. Masumura, "Evolution of Speech Recognition System—VoiceRex," NTT Technical Review, Vol. 17, No. 9, pp. 5–8, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201909fa2.html
[6] K. Nishida, I. Saito, A. Otsuka, K. Nishida, N. Nomoto, and H. Asano, "Toward Natural Language Understanding by Machine Reading Comprehension," NTT Technical Review, Vol. 17, No. 9, pp. 9–14, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201909fa3.html
[7] T. Hasegawa, Y. Sekiguchi, S. Yamada, and M. Tamoto, "Automatic Knowledge Assistance System Supporting Operator Responses," NTT Technical Review, Vol. 17, No. 9, pp. 15–18, 2019.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201909fa4.html

Kimihito Tanaka
Senior Research Engineer, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a BIS and M.E. from the University of Tsukuba, Ibaraki, and an MBA from the University of Manchester, UK. He joined NTT Human Interface Laboratories in 1995 and studied speech synthesis technologies. His current research interests are innovation management and AI technologies.

Takashi Yagi
Senior Research Engineer, Supervisor, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.E. in electrical engineering in 1990 and an M.E. in computer science in 1992 from Keio University, Tokyo. He joined NTT Human Interface Laboratories in 1992. His research interests include human-computer interaction, computer-mediated communication, and AI. He is a member of ACM (Association for Computing Machinery), IEICE (Institute of Electronics, Information and Communication Engineers), IPSJ (Information Processing Society of Japan), and VRSJ (Virtual Reality Society of Japan).

Tetsuya Iizuka
Vice President, Head of NTT Media Intelligence Laboratories.
He received a B.E. and M.E. in information engineering from Gunma University in 1989 and 1991. He joined NTT in 1991 and conducted research on database systems and data mining. He was also involved in the development of the SIP (session initiation protocol) server of Hikari Phone, development and support of open source software, personnel management, and branch office management.

Evolution of Speech Recognition System—VoiceRex

Takanobu Oba, Tomohiro Tanaka, and Ryo Masumura

Abstract

Speech recognition is a key element of artificial intelligence for contact centers. It is now used in a wide range of scenarios, supporting business in various ways. Research and development of speech recognition has a long history and has been built upon various technologies to reach today's standards. We introduce the VoiceRex speech recognition system developed by NTT Media Intelligence Laboratories, its history, and some technologies employed in the latest VoiceRex system, which are much anticipated for use in contact centers.

Keywords: speech recognition, VoiceRex, contact center

1. History of VoiceRex

Speech recognition is a key technology to understand human communication and is a necessary element of artificial intelligence (AI) for contact centers. Speech recognition is technology to convert speech in an input signal into text. Research and development (R&D) of speech recognition at the NTT laboratories has a long history, spanning half a century. NTT Media Intelligence Laboratories has developed the VoiceRex speech recognition system based on the results of these long years of research and is providing the technology to NTT Group companies, where it will be applied in a wide range of service fields.

The idea of using speech recognition to analyze contact center calls was in mind from the initial stages of this R&D. This capability was almost unthinkable at the time, but it was set as a goal to be reached some decades in the future. VoiceRex (or its predecessor speech recognition library) was first released in the early 1990s, but at that time, it was only able to recognize keywords. The current ability to recognize longer utterances and conversations was achieved in 2000. Even so, it was not at a level where it could correctly recognize conversations between two people. It could only recognize formal language such as newspaper text if it was read out clearly, and the recognizable vocabulary was very limited.

Then VoiceRex went through several technical innovations, and the performance increased dramatically. In 2008, we incorporated a weighted finite state transducer for the first time in Japan. This enabled it to learn approximately 100 times more words than before and to recognize speech from among roughly 10 million words. In 2009, this advancement was used in a system that creates a record of debates in the Japanese House of Representatives. In conditions where one person speaks at a time in question-answer format, the system was able to achieve 90% recognition accuracy. This was the beginning of the use of speech recognition to replace manual shorthand in the main assembly and various committees.

Thereafter, speech databases continued to be extended and consolidated, computing performance improved, and technology was created to utilize large-scale databases efficiently, so the performance of speech recognition continued to increase steadily. Finally, speech recognition for contact center calls started to become practical, and in 2014, NTT Software released its ForeSight Voice Mining* product for contact centers.

For several years before that, another technology had been attracting attention in the speech recognition

* ForeSight Voice Mining is currently distributed by NTT TechnoCross.

research community. Deep neural networks had emerged on the scene. Deep learning caused a great paradigm shift in speech recognition. The performance of conversion from sound signals (air pressure fluctuations) to an acoustic model (sequences of phonemes, which are vowel and consonant sounds) increased sharply, which led to dramatic increases in recognition rates for phone calls.

A commercial version of VoiceRex using deep learning was released in 2014. Then, in 2015, we used a type of neural network called CNN-NIN (convolutional neural network - network in network) to perform speech recognition on sound from a mobile terminal in a noisy public area. This resulted in a first place award among participating research institutions at the CHiME3 (The 3rd CHiME Speech Separation and Recognition Challenge) international technology evaluation event. With the spread of smartphones, people are making telephone calls more frequently while they are out and about, and we are now able to recognize speech accurately in audio signals containing more ambient noise, as when calling from a crowded location.

Through such technical innovations, applications of speech recognition have expanded rapidly. VoiceRex is now used in many products and services. In particular, the number of installed AI-related contact center products have increased rapidly, and these have become a core segment of speech recognition products and services.

However, some new issues have arisen as more and more customers have started using the technology. One such issue concerns the diversity in terms of topics. Current speech recognition technology can work more accurately if the topics in the input call audio are known ahead of time. For example, the names of services being handled will differ for each company operating a call center. Even within one company, the departments for registering or canceling membership, for receiving complaints, and for answering technical questions are all separate. Thus, current technology tunes language models for content separately for each company and also for each contact center.

Another issue is handling the flow of conversation. When two people talk, they often do not speak in clear statements. For example, when a sentence ends in "…-tion," which would be written "tʃən" using a phonetic pronunciation syllable, it might only be pronounced "tʃ," and the remaining "ən," while pronounced, may not be clear and may only be expressed in the rhythm of the speech. This can occur frequently. For such cases of unclear or omitted pronunciation, a mechanism to predict words from the context is needed. Below, we introduce a new technology incorporated in the latest version of VoiceRex to overcome this issue.

2. Conversational context language model

A language model is a model that predicts the sequence of individual words. Intuitively, it has the role of deciding whether a sentence is correct as a language or not. Our speech recognition system uses a probabilistic model called an N-gram language model, which creates models based on the idea that a word is dependent on the N-1 preceding words. The number of word combinations increases exponentially as N increases, so at most, three or four N values are used. As such, models only consider very local context. For short utterances, an N-gram language model is sufficient, but as utterances get longer, it becomes necessary to consider more context.

Thus, language models using neural networks have attracted attention recently, particularly recurrent neural network (RNN) language models, which are able to handle context over longer periods. When these models are used for speech recognition, multiple candidate speech recognition results are first obtained, and each result is scored using an RNN language model to determine the final recognition result. This is called the rescoring method.

One issue with RNN language models is that they are limited to using context in a single utterance. For calls to a contact center, using context that spans utterances is very important. For example, when the operator answers a customer's question, the content of the answer must be related to the content of the customer's question. Therefore, we have developed conversational context language model technology [1] that considers context over longer periods, spanning utterances. When it performs speech recognition, it uses the speech recognition results from successive utterances as context. In VoiceRex, a conversational context language model is applied using the rescoring method for each utterance to select better (closer to correct) sentences, and these results are used as context to perform speech recognition on the next utterance. In this way, the effectiveness of the conversational context language model is gained with the input of each utterance.

3. Neural error corrective language model

As stated before, as fluency increases, cases of

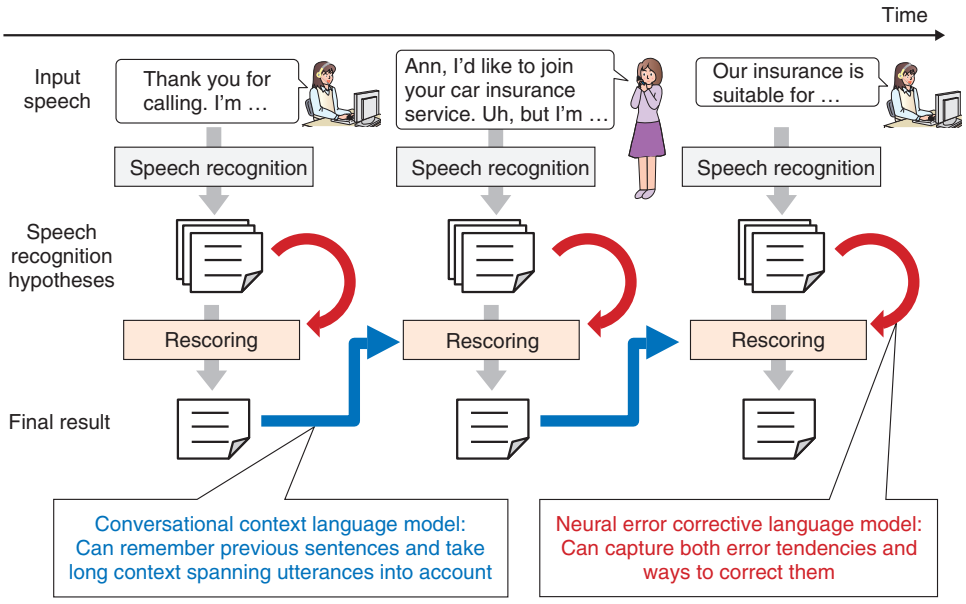


Fig. 1. Flow of applying conversional context language model and neural error corrective language model.

unclear pronunciation appear, but certain tendencies also appear in such cases. This frequently occurs for sentences ending in "-tion" (tʃən) as mentioned earlier, and also for prepositions and frequently occurring expressions such as "Thank you very much." Speech recognition tends to make the same mistakes in each of these cases. There are also many other error patterns that occur frequently, beyond those caused by unclear pronunciation. The approach of detecting such error tendencies and correcting them is called error correction.

We were able to improve recognition accuracy by developing neural error corrective language model technology [2] that incorporates a framework for considering speech recognition errors into an RNN language model. Specifically, we introduced a neural network called an encoder-decoder model, which provides a mechanism to select the correct sentence from among results that include speech recognition errors. To train the neural error corrective language model, we used sentences that produced speech recognition errors together with the corresponding correct sentences, and trained for the relationship between them. In this way, we captured both the error tendencies and ways to correct them.

When performing speech recognition, we use the rescoring method, in the same way as for the conversational context language model. This model can be used together with the conversational context language model, producing results that are augmented by both language models for each utterance. This is illustrated in **Fig. 1**. During a call, utterances are extracted from the input signal, multiple recognition result candidates are output for each, and both models are applied to select the final recognition result. For the conversational context language model, this final recognition result is returned as context for recognition of the next utterance.

4. Future prospects

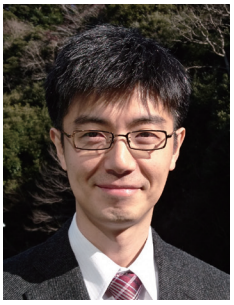
For most of the contact centers considered so far, the customers would initiate the call to the contact center. In such cases, the conversation between the operator and customer is one-time-only and relatively formal. This is good for the accuracy of speech recognition. However, as AI products are introduced into contact centers, they are being used more and more by companies to contact their customers. In such cases, an operator is assigned to each customer, and the operator speaks to the customer multiple times to encourage more frank conversation, which complicates speech recognition. As in the past, as the range of applications has broadened, situations are encountered that make recognition more difficult than before. Each time this occurs, R&D is conducted to overcome the difficulty, and the technology continues to advance. We will continue to advance VoiceRex,

meeting the new challenges encountered by users as they use speech recognition in real life.

References

[1] R. Masumura, T. Tanaka, A. Ando, H. Masataki, and Y. Aono, "Role Play Dialogue Aware Language Models Based on Conditional Hierarchical Recurrent Encoder-Decoder," Proc. of the 19th Annual Conference of the International Speech Communication Association (Interspeech 2018), pp. 1259–1263, Hyderabad, India, Sept. 2018.

[2] T. Tanaka, R. Masumura, H. Masataki, and Y. Aono, "Neural Error Corrective Language Models for Automatic Speech Recognition," Proc. of Interspeech 2018, pp. 401–405, Hyderabad, India, Sept. 2018.



Takanobu Oba

Senior Research Engineer, Audio, Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E. and M.E. from Tohoku University, Miyagi, in 2002 and 2004, and a Ph.D.(Eng.) from Tohoku University in 2011. He joined NTT Communication Science Laboratories in 2004, where he has been engaged in research on spoken language processing. He received the 25th Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2008. He joined NTT DOCOMO in 2015 and worked on service development of spoken dialogue systems such as docomo AI Agent. He is currently working on contact center related solutions at NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ.

Tomohiro Tanaka

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E. from Tokyo University of Science in 2015 and an M.E. from Tokyo Institute of Technology in 2017. Since joining NTT in 2017, he has been researching automatic speech recognition and spoken language processing. He was the recipient of the 13th Best Student Presentation Award of ASJ. He is a member of ASJ.

Ryo Masumura

Distinguished Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received a B.E., M.E., and Ph.D. in engineering from Tohoku University, Miyagi, in 2009, 2011, and 2016. Since joining NTT in 2011, he has been researching speech recognition, spoken language processing, and natural language processing. He received the Student Award and the Awaya Kiyoshi Science Promotion Award from ASJ in 2011 and 2013, respectively, the Sendai Section Student Awards Best Paper Prize from the Institute of Electrical and Electronics Engineers (IEEE) in 2011, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2014, the Young Researcher Award from the Association for Natural Language Processing (NLP) in 2015, and the ISS Young Researcher's Award in Speech Field from IEICE in 2015. He is a member of ASJ, IPSJ, NLP, IEEE, and the International Speech Communication Association.

Toward Natural Language Understanding by Machine Reading Comprehension

Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Kosuke Nishida, Narichika Nomoto, and Hisako Asano

Abstract

The ability of artificial intelligence (AI) to comprehend text is becoming a major topic of discussion. While natural language understanding by AI poses a difficult problem, a significant improvement in AI reading comprehension has been achieved in recent years through the application of deep learning techniques. This article introduces machine reading comprehension technology now under research and development at NTT with the aim of achieving an agent that can understand business knowledge written in manuals, understand the language used by customers, and provide appropriate answers to questions.

Keywords: machine reading comprehension, natural language understanding, deep learning

1. Machine reading comprehension and natural language understanding

NTT Media Intelligence Laboratories is undertaking the research and development (R&D) of machine reading comprehension to support customer reception at contact centers through artificial intelligence (AI). Machine reading comprehension is a technology involving the use of AI to read the content of manuals, contracts, and other documents and reply to questions. It takes up the challenge of achieving *natural language understanding*, that is, the understanding of everyday human language. The aim here is to support operators at a contact center by having AI interpret the content in manuals and find exact answers even if FAQs (frequently asked questions) have not been prepared beforehand from manuals for search purposes (**Fig. 1**).

Machine reading comprehension is a new research field that is developing rapidly thanks to progress in deep learning and the preparation of large-scale datasets. It attracted much attention in January 2018 when AI achieved a higher score than humans on the Stanford Question Answering Dataset (SQuAD) [1]—a machine reading comprehension dataset—prepared by Stanford University. However, the problem setup in SQuAD is relatively simple, and for more difficult problem setups, AI still comes up short against human reading comprehension. While continuing to refine machine reading comprehension technology amid academic competition using datasets for research purposes, researchers at NTT Media Intelligence Laboratories are working to solve key problems with the aim of providing a practical system for contact centers. This article introduces our research results to date in machine reading comprehension.

2. Large-scale machine reading comprehension technology for finding answers from many documents

Problem setups in early research of machine reading comprehension limited the knowledge source to the text in a single document, but actual application scenarios including contact centers require that answers to questions be found from many documents. However, comprehending in detail the content of all such documents based on a machine comprehension

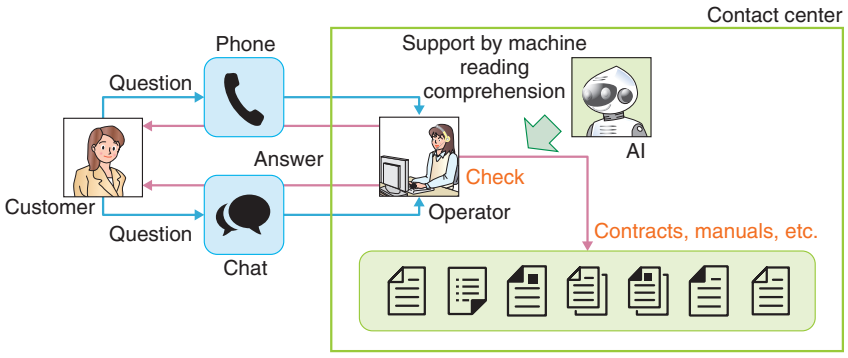


Fig. 1. Contact center and machine reading comprehension.

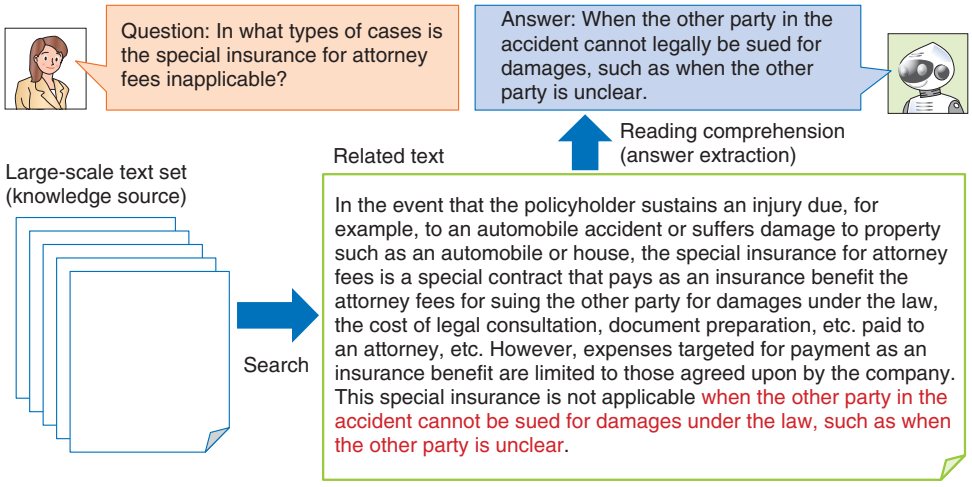


Fig. 2. Large-scale machine reading comprehension.

model would slow down system operation, so there is a need for narrowing down the documents needed for finding an answer in a high-speed and accurate manner.

In response to this need, NTT Media Intelligence Laboratories established large-scale machine reading comprehension technology [2]. This technology first narrows down the documents related to the input question in a broad manner using high-speed keyword search technology. It then further narrows down related documents using a neural search model, and finally, it finds an answer using a neural reading comprehension model (**Fig. 2**). Thus, we have greatly improved information retrieval accuracy by having AI learn neural information retrieval and neural reading comprehension simultaneously with one model. As a result, we achieved the world's highest question answering accuracy [2] in a machine reading comprehension task against a set of five million Wikipedia articles in English.

3. Explainable machine reading comprehension technology for understanding and presenting evidence written at multiple locations

It is difficult to guarantee that answers will be 100% correct by AI based on machine learning, so it is important to have a function that enables humans to check the validity of an answer that is output by a machine reading comprehension model. A machine reading comprehension model that can present where in the document set information that serves as evidence for an answer is written would therefore be useful. However, there are cases in which information

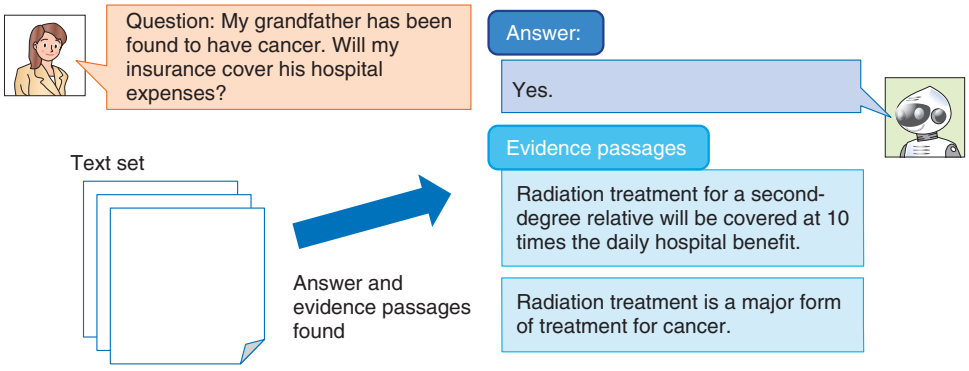


Fig. 3. Explainable machine reading comprehension.

that provides evidence for an answer is divided among multiple passages within the document set. Understanding and extracting all instances of such information is known to be a formidable problem [3].

At NTT Media Intelligence Laboratories, we have addressed this problem by establishing explainable machine reading comprehension technology that combines a model inspired by document summarization technology with a machine reading comprehension model (**Fig. 3**) [4]. Document summarization technology extracts important sentences within text as a summary, but our proposed technique extracts sentences that are important for the answer as evidence sentences. Extracting evidence sentences using a neural evidence extraction model while simultaneously finding answers with a neural reading comprehension model has enabled us to present both an answer and evidence with good accuracy even for difficult problems in which the evidence is divided among multiple passages. Consequently, on applying this technology to the HotpotQA [3] task of outputting an answer and its evidence with respect to questions made against Wikipedia articles in English, we were able to take first place on the leaderboard.

4. Specific query generation technology to clarify a vague question

Datasets for research use in the field of machine reading comprehension are commonly prepared in such a way that an answer can be uniquely specified. However, in an actual contact center or similar operation, there are cases in which the intent of the question input by a customer is vague, preventing an answer from being found. In such situations, more natural communication could be achieved if AI could appropriately ask the customer about the intent of the question much like a human operator does.

With this in mind, NTT Media Intelligence Laboratories established specific query generation technology to rewrite a vague question into specific questions that can be answered by machine reading comprehension (**Fig. 4**) [5]. The proposed technique extracts candidate answers based on machine reading comprehension with respect to the input question and generates revised questions that remove question ambiguity for each candidate answer.

The example in Fig. 4 is on cancellation fees. As another example, we take the question "What is the maximum insurance benefit provided by special insurance for attorney fees?" In specific query generation, the technique generates multiple specific queries instead of immediately presenting an answer, such as "What is the maximum insurance benefit provided by special insurance for attorney fees for one accident?" and "What is the maximum insurance benefit provided by special insurance for attorney fees for legal consultation and document preparation?" The customer is then asked to select the question nearest to the information desired. In this example, the answer when selecting the former question is "3,000,000 yen," while the answer when selecting the latter question is "100,000 yen." This technique enables machine reading comprehension to be applied to a broader range of questions.

5. Generative machine reading comprehension technology for generating multi-style answers

In machine reading comprehension, the most common type of problem setup involves extracting an answer from text serving as a knowledge source.

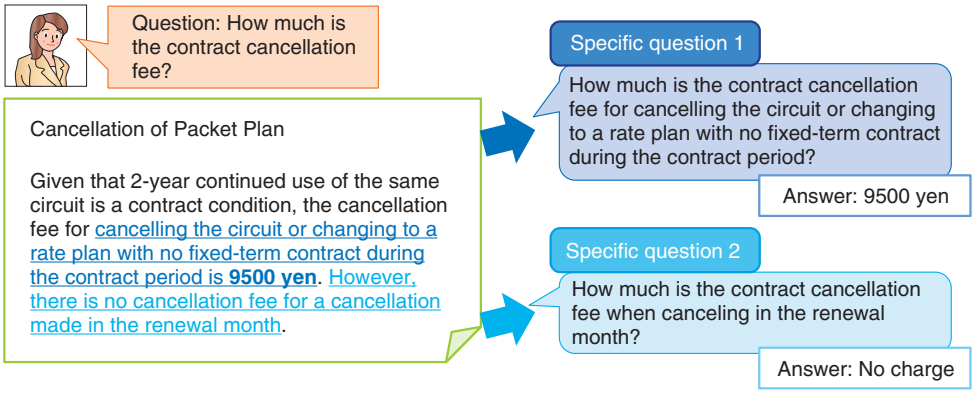


Fig. 4. Specific query generation technology.

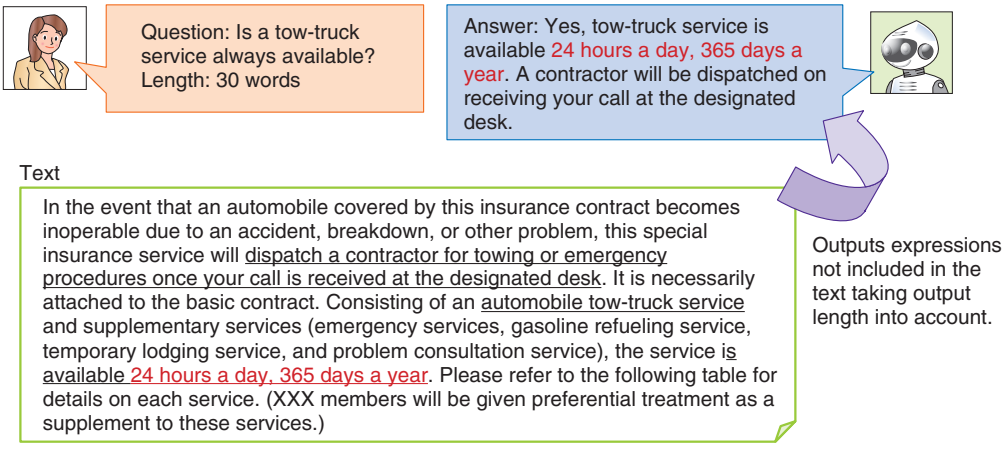


Fig. 5. Generative machine reading comprehension and answer summarization.

Nevertheless, there is a need to generate a more natural style of writing in the case of smart devices, chatbots, and other types of advancing technology. However, the degree of difficulty in generating such an answer is high, and the amount of training data is still insufficient, so worldwide research in this area has yet to advance.

In response to this problem, NTT Media Intelligence Laboratories established generative machine reading comprehension technology that can generate answers with expressions not found in the source text (**Fig. 5**) [6]. For example, given the question "When is a tow-truck service available?" the proposed technique can generate a natural-sounding answer in the manner of "Tow-truck service is available 24 hours a day, 365 days a year," which includes not just source text but also appropriate content from the question itself. This technique can simultaneously learn from machine reading comprehension data with different answering styles and enables the answering style to be selected at the time of generation. This alleviates the problem of having insufficient training data. We applied this technique to two tasks with different answering styles against the MS MARCO [7] dataset that performs open-domain question answering using actual search engine logs and were able to take first place on the leaderboard in that task.

6. Answer summarization technology for controlling answer length

For a chatbot that replies to questions from customers, presenting long answers output from a machine reading comprehension model in their original form

makes for difficult reading by customers. There is therefore a need for appropriately adjusting the length of an answer. In addition, adjusting the length of answers according to present conditions, such as when reading answers on a smartphone or when reading on a personal computer, enables flexible answering tailored to the customer or the current device.

In recognition of this need, NTT Media Intelligence Laboratories established answer summarization technology for controlling answer length using a neural network (Fig. 5). This technology achieves a function for appropriately summarizing an answer to a question by combining a neural network model that identifies important words in text and another neural network model that summarizes (generates) an answer from the words. Here, giving information on length in embedding vector form when summarizing an answer enables that answer to be output in any specified length.

7. Future development

Going forward, we plan to identify problems hindering practical application of machine reading comprehension technology to question answering based on manuals in a contact center. We also plan to feed these problems back to R&D to make technical improvements with the aim of supporting operators through AI and achieving automatic answering. Machine reading comprehension—the understanding of natural language used by humans—is a difficult challenge, but it should foster innovation in various NTT Group agent-based AI services for contact centers and beyond. NTT Media Intelligence Laboratories is committed to further R&D in this field with the aim of achieving AI that can communicate with human beings using natural language.

References

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), pp. 2383–2392, Austin, Tx, USA, Nov. 2016.

[2] K. Nishida, I. Saito, A. Otsuka, H. Asano, and J. Tomita, "Retrieve-and-Read: Multi-task Learning of Information Retrieval and Reading Comprehension," Proc. of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), pp. 647–656, Torino, Italy, Oct. 2018.

[3] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering," Proc. of EMNLP 2018, pp. 2369–2380, Brussels, Belgium, Oct./Nov. 2018.

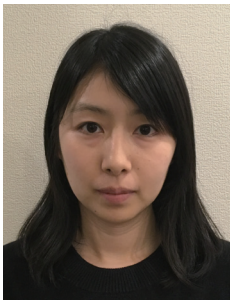
[4] K. Nishida, K. Nishida, M. Nagata, A. Otsuka, I. Saito, H. Asano, and J. Tomita, "Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction," The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, Jul./Aug. 2019.

[5] A. Otsuka, K. Nishida, I. Saito, H. Asano, and J. Tomita, "Specific Question Generation for Reading Comprehension," AAAI 2019 Reasoning for Complex Question Answering Workshop, Honolulu, HI, USA, Jan./Feb. 2019.

[6] K. Nishida, I. Saito, K. Nishida, K. Shinoda, A. Otsuka, H. Asano, and J. Tomita, "Multi-style Generative Reading Comprehension," ACL 2019, Florence, Italy, Jul./Aug. 2019.

[7] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," arXiv, 1611.09268v3, 2018.

Kyosuke Nishida
Distinguished Researcher, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.E., MIS, and Ph.D. in information science and technology from Hokkaido University in 2004, 2006, and 2008. He joined NTT in 2009. He was a chief engineer at NTT Resonant from 2013 to 2015. He received the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2015 and the Kambayashi Young Researcher Award from the Database Society of Japan (DBSJ) in 2017. His current interests include natural language processing, artificial intelligence, and web and spatial data mining. He is a member of the Association for Computing Machinery (ACM), IPSJ, the Association for Natural Language Processing (NLP), the Institute of Electronics, Information and Communication Engineers (IEICE), and DBSJ.



Itsumi Saito
Researcher, Knowledge Media Project, NTT Media Intelligence Laboratories.
She received a B.E. from Waseda University, Tokyo, in 2010 and an M.S. from the University of Tokyo in 2012. She joined NTT in 2012. Her research interests include natural language processing. She is a member of IPSJ and NLP.

Atsushi Otsuka
Researcher, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.S. and M.S. in information science from University of Tsukuba, Ibaraki, in 2011 and 2013. He joined NTT Media Intelligence Laboratories in 2013. His research interests include natural language processing, particularly dialogue processing, and information retrieval. He is a member of IPSJ, the Japanese Society for Artificial Intelligence, and DBSJ.

Kosuke Nishida
Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.E. in engineering and MIS in information science and technology from the University of Tokyo in 2015 and 2017. He joined NTT in 2017. His research interests include natural language processing and mathematical informatics. He is a member of the Association for Computational Linguistics and NLP.

Narichika Nomoto
Senior Research Engineer, Knowledge Media Project, NTT Media Intelligence Laboratories.
He received a B.A. and M.M.G. in information science and technology from Keio University, Tokyo, in 2005 and 2007. He joined NTT in 2007 and studied speech and emotion recognition. He worked as a data scientist at NTT Communications from 2012 to 2016. He is currently engaged in R&D of natural language processing at NTT Media Intelligence Laboratories.

Hisako Asano
Senior Research Engineer, Supervisor, Knowledge Media Project, NTT Media Intelligence Laboratories.
She received a B.E. in information engineering from Yokohama National University, Kanagawa, in 1991. She joined NTT Information Processing Laboratories in 1991. Her research interests include natural language processing, especially morphological analysis, and information extraction. She is a member of IPSJ and NLP.

Automatic Knowledge Assistance System Supporting Operator Responses

Takaaki Hasegawa, Yuichiro Sekiguchi, Setsuo Yamada, and Masafumi Tamoto

Abstract

The variety and complexity of products and services handled by contact centers has increased recently, which places a heavy load on operators, as they must retain more information. This has led to decreasing operator retention rates. This article introduces an automatic knowledge assistance system that assists operators by automatically presenting appropriate information (knowledge) to them while they are handling calls.

Keywords: contact center AI, dialogue understanding, FAQ search

1. Introduction

Contact centers have the important role of being the point of contact between an enterprise and its customers. A major objective of a contact center is to increase customer satisfaction by responding to inquiries quickly and appropriately. However, the variety and complexity of products and services being handled continues to increase, requiring operators to know an increasing amount of information. This places a heavy burden on operators and has resulted in decreasing operator retention rates.

We are developing an automatic knowledge assistance system that supports operators by automatically presenting appropriate information (knowledge) to them as they respond to calls, particularly if they have little experience in the business. However, creating and maintaining the information to be presented to operators can be expensive. To reduce this cost, we are also working on technology to assist with consolidation of frequently asked questions (FAQs), which are one form of such information. This article introduces the automatic knowledge assistance system and the FAQ consolidation technology.

2. Overview of automatic knowledge assistance system

The automatic knowledge assistance system provides support to operators who are less experienced in the business while they are handling calls by presenting appropriate information based on the issue raised by the contact center (or call center) caller. Text of the ongoing dialogue between the operator and customer is shown on the left of the screen viewed by the operator, and several similar high-scoring questions with answers are displayed on the right, which are found automatically in the FAQs, based on statements by the customer regarding the issue and the operator's responses to confirm the issue.

The system performs the following steps:

(1) Conversion of the dialogue to text: Speech recognition is applied to the dialogue between operator and customer, and utterance segmentation is applied to the results so they can be presented in a textual form that is easy for the operator to read.

(2) Dialogue structuring: Dialogue segmentation is used to apply structure to the dialogue,

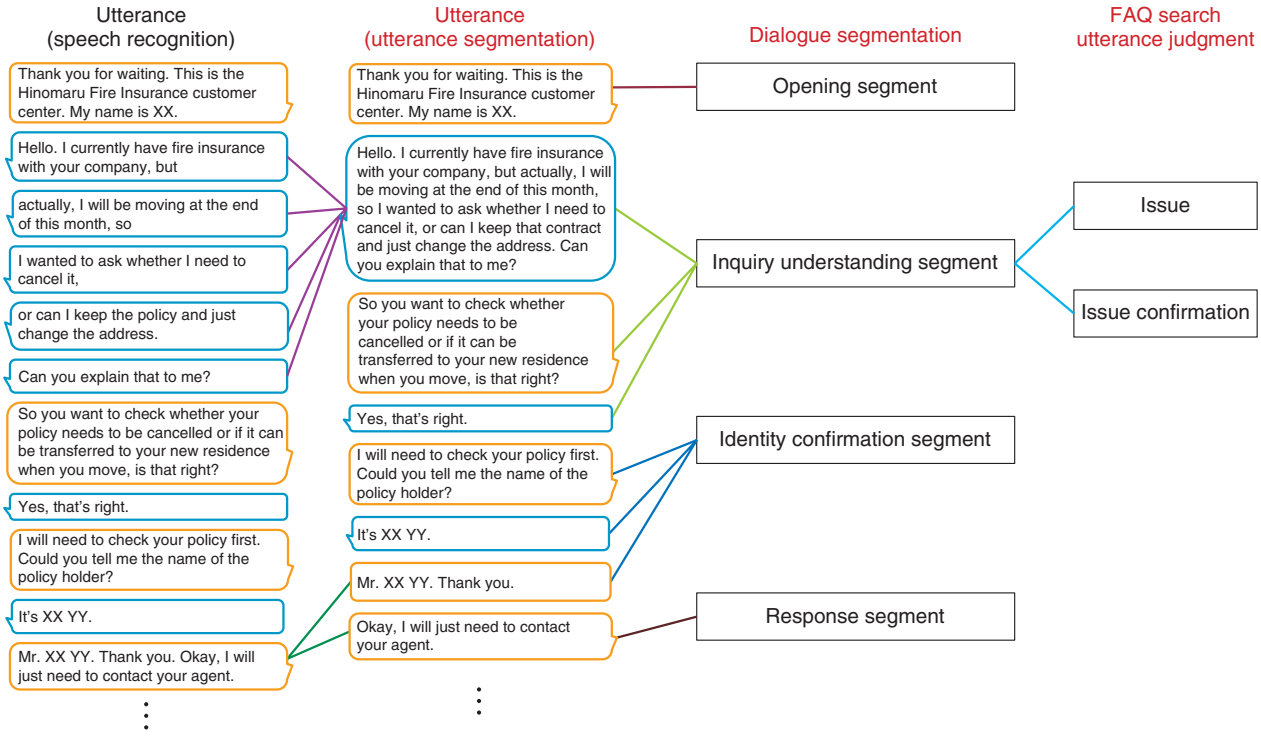


Fig. 1. Automatic knowledge assistance steps.

based on features found in contact center interactions.

(3) FAQ search automation: An FAQ search utterance judgment function selects customer utterances that represent the customer's issue and operator utterances to confirm the issue. These are used to search automatically for previously created FAQ issues.

The processes from speech recognition through utterance judgment are illustrated in **Fig. 1**.

2.1 Conversion of speech dialogue to text

Speech recognition technology is used to convert the speech to text, but it is difficult for the operator to see the interaction between the operator and the customer if the results of speech recognition are simply displayed as-is. When the customer calls, they are thinking about the issue as they describe it and may speak slowly, pause, or stutter. As-is, the speech recognition results from the customer's utterances could be divided up with punctuation where there was just a pause, when they should really be displayed as a semantic unit. Thus, displaying them as-is can make the results difficult to understand.

We implemented a function called utterance segmentation that displays utterances in relatively longer units. This function receives the results from speech recognition and determines whether the operator or customer has finished talking, and whether they have made a definitive statement. The decision is made using a deep neural network (DNN) trained using data of approximately 1000 conversation logs. This is complemented with a heuristic that determines whether the speaker has finished by finding points where the dialogue switches from one speaker to the other, while also ignoring confirming sounds made by the other party ("Oh," "uh huh," "Got it," etc.). Overall, this achieves very accurate results.

2.2 Structuring speech dialogue

Dialogues with contact centers tend to be centered on tasks related to the business such as inquiries regarding products or requests for procedures related to a particular service, so typical patterns appear in the flow of dialogue. For example, at a call center receiving inquiries for an insurance product, operators begin by introducing themselves, checking the reason for the call, and confirming the policy holder and policy details before handling the incident. Then they would conclude the call with some kind of

formality. Having an overall grasp of a conversation flow in this way provides strong clues for understanding the dialogue between operator and customer.

We call the function that determines this conversation flow dialogue segmentation. We designed labels suitable for response segments at an inbound call center, created training data of approximately 1000 call center conversation logs, and trained a DNN model to implement technology able to predict dialogue segments accurately.

2.3 Automating FAQ search

To search the FAQs automatically, two conditions must be satisfied. An appropriate query for the search must be selected, and the timing of the search must be appropriate. For the search query, the speech recognition results are not used as-is. The results of utterance segmentation enable us to avoid searching with statement fragments and instead perform queries for full utterances. In the example in Fig. 1, making a query using the partial utterance, "…actually, I will be moving at the end of this month, so…" might result in a search with just the keyword, "moving." Without the keywords that follow such as "cancel" or "change the address," FAQ entries useful to the operator would not be found. Use of utterance segmentation enables the system to select a query that is appropriate for the FAQ search.

Performing a search every time the customer or operator makes an utterance is also too frequent and makes it difficult for the operator to check the search results, so it is necessary to determine which utterances represent the issue. We used machine learning to implement local classifiers, including issue utterance judgment, which determines whether a customer utterance represents an issue, and issue confirmation utterance judgment, which determines whether an operator utterance is confirming the issue. By combining these with the dialogue segmentation described earlier, issue utterance judgment and issue confirmation utterance judgment are applied only to utterances in the inquiry understanding segment. This prevents the selection of irrelevant issues based on utterances in the identity confirmation segment, for example, which in turn increases the accuracy of issue selection. Through this process, queries are made with keywords selected from utterances that have been determined to be related to the issue, and FAQ searches are implemented with appropriate timing.

3. Assisting FAQ creation

Even with FAQ searches based on issue utterances and issue confirmation utterances, if the FAQs being searched are not well organized, a suitable search result cannot be returned. FAQs are composed of question and answer sentences, and since the risk is high if an operator provides an incorrect response, answer sentences must be created and organized by a contact center supervisor that is knowledgeable about the business being handled by the center. However, questions must be created based on details that customers actually query, so they must be regarding matters that are actually common in inquiries. We implemented an FAQ consolidation assistance technology that analyzes past conversation logs accumulated by the contact center, selects candidates for FAQs, and presents them to a contact center supervisor (**Fig. 2**).

Partial utterances that could represent issues are first extracted from the conversation logs using the issue utterance judgment function described above. Issues that customers ask about are many and varied, but issues that have been asked several times in the past are candidates to be added to the FAQs. Thus, issues are collected and filtered according to their frequency of occurrence. In this process, words and phrases that give clues of the issue are reduced to a multi-dimensional vector and used to cluster similar issues. Then, issues that are larger than a set scale are selected from the summarized issue sets and presented as question candidates.

In the past, when building automatic knowledge assistance systems, workers needed to search large conversation logs for question candidates and send them to the center supervisor. Automating this part of the process will contribute to reducing the cost of implementing such systems.

4. Future development

This article introduced technology that extracts issues from dialogues between operators and customers at a contact center and uses them to search FAQs. Much of the functionality introduced is implemented using machine learning, so important outstanding issues include creating large amounts of training data at low cost each time this system is introduced into a contact center, and the ability to generalize models so that they can be applied to various tasks. In the future, we will continue to work on increasing the accuracy of FAQs, and also on reducing the cost of deploying the system.

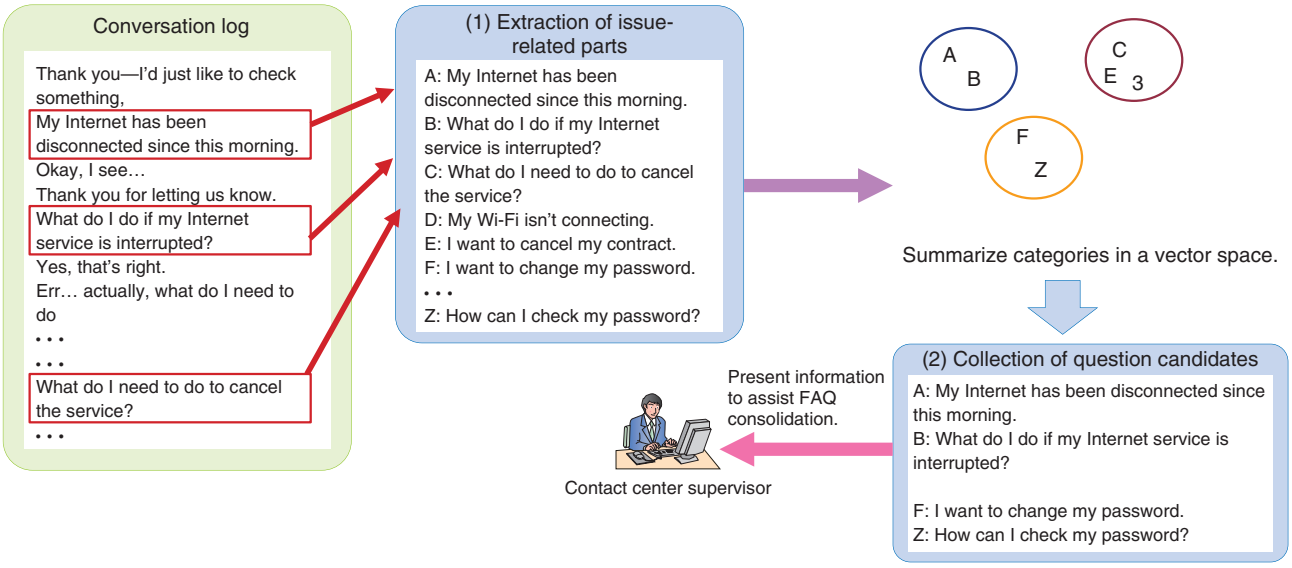


Fig. 2. FAQ consolidation assistance technology.

Takaaki Hasegawa
Senior Research Engineer, Social Knowledge Processing Laboratory, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from Keio University, Tokyo, in 1992 and 1994 and a Dr. Eng. from Tokyo Institute of Technology in 2010. Since joining NTT in 1994, he has been engaged in the research of natural language processing and intelligent information access. He was a visiting researcher at New York University from 2003 to 2004. He is a member of the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence, and the Association for Natural Language Processing (NLP).

Yuichiro Sekiguchi
Senior Research Engineer, Social Knowledge Processing Laboratory, NTT Media Intelligence Laboratories.
He received a B.S. and M.S. in mechanoinformatics from the University of Tokyo in 2002 and 2004. He joined NTT in 2004 and has been researching information retrieval and natural language processing. He is a member of IPSJ and the Database Society of Japan.

Setsuo Yamada
Research Engineer, Social Knowledge Processing Laboratory, NTT Media Intelligence Laboratories.
He received an M.S. in information science from Tokyo Denki University in 1992 and a Ph.D. in engineering from Tottori University in 2006. He joined NTT in 1992. From 1997 to 2000, he was with ATR Spoken Language Translation Research Laboratories. He rejoined NTT in 2000. His research interests include natural language processing, artificial intelligence, and software engineering. He is a member of IPSJ and NLP.

Masafumi Tamoto
Senior Research Engineer, Social Knowledge Processing Laboratory, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from Tokyo Institute of Technology in 1991 and 1993. Since joining NTT in 1993, he has been working on the research and development of dialogue understanding systems and knowledge media technologies. He received the Awaya Prize Young Researcher Award from the Acoustic Society of Japan (ASJ) in 1999. He is a member of IPSJ and ASJ.

Towards Secured and Transparent Artificial Intelligence Technologies in Hierarchical Computing Networks

Yitu Wang and Takayuki Nakachi

Abstract

Researchers at NTT Network Innovation Laboratories have recently been focusing on the interdiscipline of transparent artificial intelligence (AI) technologies and hierarchical computing networks. A hierarchically distributed computing structure not only improves the quality of computation but also creates an extra degree of diversity for algorithm refinement. Sparse coding, an important transparent AI technique, is finding application in this new domain. We propose in this article a secure sparse coding scheme that enables computing directly on cipher-texts. We also demonstrate its application to image compression and face recognition in edge and cloud networks.

Keywords: secure sparse coding, edge and cloud, image compression, face recognition

1. Introduction

We are witnessing a sense of excitement in the research community and a frenzy in the media regarding advances in artificial intelligence (AI). Remarkable progress has been made in a variety of AI tasks such as image classification and speech comprehension by making use of deep neural networks [1]. However, we need to be able to fully trust the algorithmic prescriptions before we can readily accept and apply them in practice. For this reason, interpretability and explainability are two essential ingredients in AI algorithm design.

Sparse coding was inspired by the sparsity mechanism of nature [2] and has received considerable attention as a representative transparent AI technique. For example, the sparsity mechanism exists in the human vision system. A learning algorithm that attempts to find sparse linear codes for natural scenes will develop a complete family of localized, oriented, bandpass receptive fields. It has proved to be an extraordinarily powerful solution in a wide range of application fields, especially in signal processing, image processing, machine learning, and computer vision [3]. In addition, sparse coding demonstrates strong potential in practical implementations because it is sufficiently flexible to capture much of the variation in real datasets and provides insights into the features extracted from the dataset.

For the purposes of practical application and providing further support to real-time services, the computational complexity involved should not be overlooked. These complex and well-engineered AI approaches require substantial effort in parameter tuning, and they pose exigent requirements on computation capability, which cannot be easily satisfied by solely relying on devices due to their limited resources, for example, limited computation capability and memory size.

A cloud built on top of a datacenter, which seamlessly integrates storage and computation, could be an ideal platform for implementing the aforementioned algorithms. It, however, faces significant challenges in data collection and service distribution over the network, given that the devices in service are globally and remotely distributed. One promising solution to address these limitations is to make use of the hierarchically distributed computing structure consisting of the edge, cloud, and devices, as shown in **Fig. 1** [4]. This configuration makes it possible to

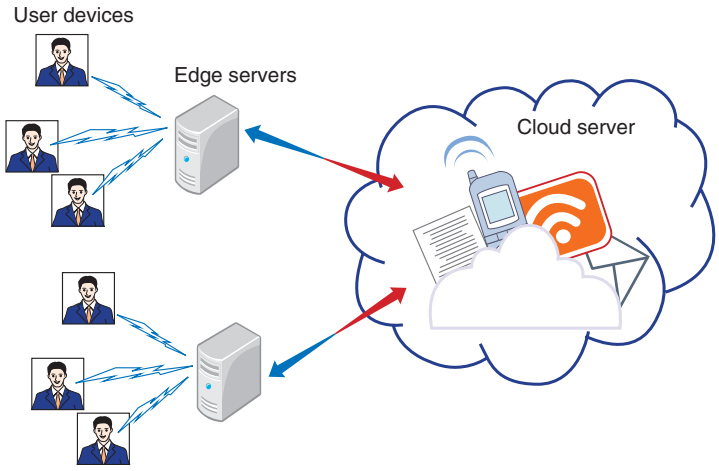


Fig. 1. Edge and cloud network.

not only substantially reduce the tension between computation-intensive applications and resource-limited devices, but also to completely avoid the long latency incurred in the information exchange between devices and the cloud in wide area networks [4]. In addition, the relative uniqueness of the information available from various devices in service prompts the algorithms to capture different patterns along the system dynamics, and in turn, creates an extra degree of diversity. Therefore, transparent AI techniques are being applied in this new domain to reduce computation demands at each device, while further enhancing the performance by exploiting the multi-device diversity.

To exploit more dimensions of edge and cloud resources for purposes other than just fulfilling computation demands, we allow the cloud to produce a joint computing result based on information obtained from each device and try to investigate what the fundamental benefit is of exploiting the multi-device diversity. However, this could lead to serious privacy concerns, as the private information could be collected and misused without permission by the third party. Commonly deployed encrypting algorithms such as advanced encryption standard and secure hash algorithm provide the capability of security, but they cannot render the designed algorithms valid; that is, computing cannot be carried out only with the encrypted data. Even though algorithms such as homomorphic encryption and secure multi-party computation enable computing on cipher-texts, they are faced with the curse of dimensionality regarding the size of data and thus incur significant computational complexity.

In this article, we report a secure sparse coding scheme with low complexity based on random unitary transform, which enables sparse modeling based algorithms to directly compute on the encrypted data. Moving one step ahead, we further demonstrate its application to image compression and face recognition in edge and cloud networks and show the superiority of the proposed framework through simulation results.

2. Secure sparse coding

As illustrated in **Fig. 2**, an observed signal set $\boldsymbol{Y}$ can be represented as the linear combination of only a few atoms from the dictionary $\boldsymbol{D}$. The core sparse representation problem is defined as the quest to find the sparsest possible representation $\boldsymbol{X}$. Due to the underdetermined nature of $\boldsymbol{D}$, this linear system offers in general many infinitely possible solutions, and among these we seek the one with the fewest nonzeros. This problem is known to be NP (nondeterministic polynomial time)-hard with a reduction to NP-complete subset selection problems in combinatorial optimization.

Alternatively, it is possible to find an approximate solution by taking the following two steps.

1) Dictionary training

Given a training set $\boldsymbol{Y}_{train}$, learning a reconstructive dictionary with K atoms for obtaining the sparsest representation can be accomplished by solving the following optimization problem,

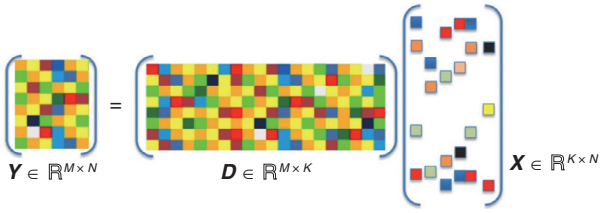


Fig. 2. Sparse coding.

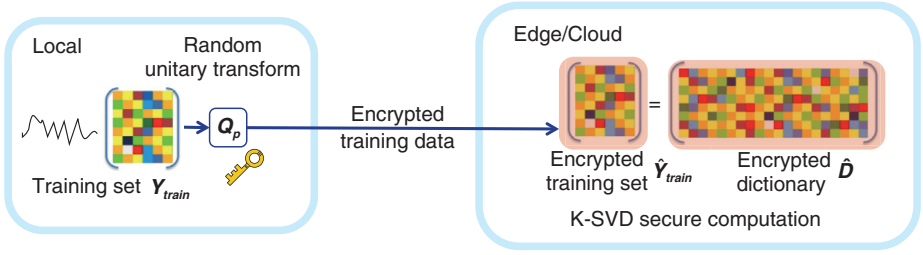


Fig. 3. Dictionary training based on encrypted data.

$$\arg\min_{D} \|\boldsymbol{Y}_{train} - \boldsymbol{D}\boldsymbol{X}\|_2^2, s.t., \|x_i\|_0 \leq \epsilon, \forall i \in \boldsymbol{Y}_{train}, \quad (1)$$

where $\boldsymbol{D}$ is the learned dictionary, $\boldsymbol{X}$ is the sparse representation, and ϵ is the sparsity constraint factor. We can apply K-SVD to find an approximate solution.

2) Sparse representation

Given a testing sample $\boldsymbol{y} \in \boldsymbol{Y}$, a sparse representation $\boldsymbol{x} \in \boldsymbol{X}$ based on the trained dictionary $\boldsymbol{D}$ can be calculated by

$$\arg\min_{x} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2, s.t., \|\boldsymbol{x}\|_0 \leq \epsilon. \quad (2)$$

The above optimization problem can be efficiently solved using orthogonal matching pursuit (OMP).

To address the security issue, we adopt random unitary transform, which not only proves to be effective for biometric template protection, but also has the desired low computational complexity for application in scenarios with a large cipher-text size [5]. Any vector v encrypted by random unitary matrix $\boldsymbol{Q}_p$ with private key p can be expressed as $\hat{\boldsymbol{v}} = \boldsymbol{Q}_p \times \boldsymbol{v}$, where $\hat{\boldsymbol{v}}$ is the encrypted vector and $\boldsymbol{Q}_p$ satisfies $\boldsymbol{Q}_p^* \times \boldsymbol{Q}_p = \boldsymbol{I}$, where $(\cdot)^*$ and $\boldsymbol{I}$ respectively represent the Hermitian transpose and the identity matrix. Gram-Schmidt orthogonalization can be adopted for generating $\boldsymbol{Q}_p$. This encryption technique has been proved to be robust in terms of brute-face attacks, diversity, and irreversibility.

The process to extract the feature dictionary from the encrypted data is depicted in **Fig. 3**. The user device first encrypts its training data locally, which are then transmitted to the edge server in close proximity via wireless channels. Then the dictionary is trained directly using the encrypted data at the edge/cloud.

We proved in a previous study [6] that the relationship between the dictionary $\hat{\boldsymbol{D}}$ trained from the encrypted data $\hat{\boldsymbol{X}}$, and $\boldsymbol{D}$ trained from the original data $\boldsymbol{X}$ satisfies $\hat{\boldsymbol{D}} = \boldsymbol{Q}_p \times \boldsymbol{D}$.

The process to obtain the secure sparse representation is shown in **Fig. 4**. The user device first encrypts its testing data locally. After receiving the encrypted data, the edge server calculates its sparse representation using the trained dictionary.

We have proved that the sparse representation $\hat{\boldsymbol{x}}$ of the encrypted data $\hat{\boldsymbol{y}}$ based on the encrypted dictionary $\hat{\boldsymbol{D}}$ is identical to $\boldsymbol{x}$ of the original data $\boldsymbol{y}$ based on the dictionary $\boldsymbol{D}$, that is, $\hat{\boldsymbol{x}} = \boldsymbol{x}$ [7].

3. Application to image compression

The traditional method for secured image transmission is based on compression-then-encryption (CtE) systems. In CtE systems, images are uploaded to social networking service (SNS) providers by users with the assumption that the entire process can be

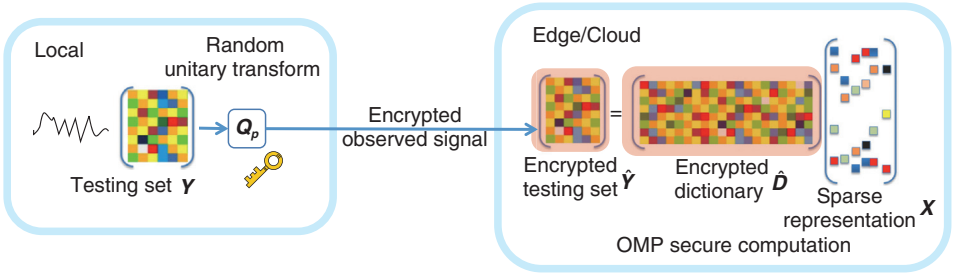


Fig. 4. Secure sparse representation.

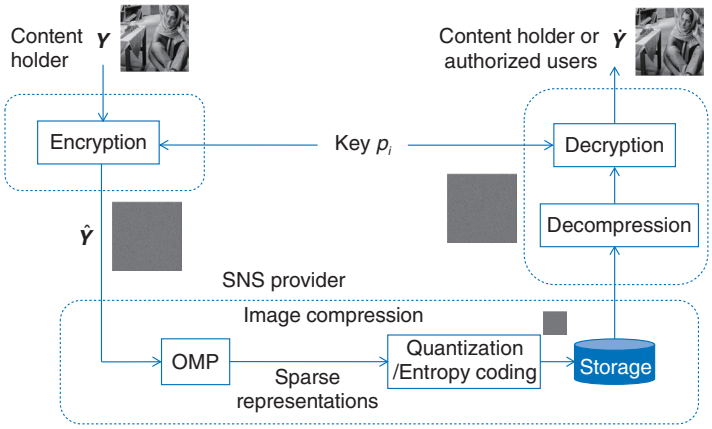


Fig. 5. EtC system using secure sparse coding.

trusted, but the privacy of the uploaded images cannot actually be controlled by the users. Therefore, there are serious concerns about the privacy of those uploaded images because SNS providers simply take full control of this process.

Encryption-then-compression (EtC) systems have been proposed to securely transmit images through an untrusted channel provider. EtC systems enable us to protect unencrypted images from the SNS providers because the encrypted images can only be recovered by authorized users, while enabling recompression by the providers. This approach supports compressing images on the cloud while keeping the image data secure.

The effectiveness of sparse coding in image compression has been reported. One study [8] showed that rate-distortion based sparse coding outperforms JPEG*1 and JPEG 2000 up to 6+ dB and 2+ dB, respectively. An EtC system using the proposed secured sparse coding for image archives and sharing in SNSs is illustrated in **Fig. 5**. We obtain the sparse representations by feeding the encrypted dictionary $\hat{D}$ and the encrypted image $\hat{Y}$ into the secure OMP computation. The proposed algorithm can easily control the number of sparse representations without decrypting the encrypted images. To this end, the rate-distortion tradeoff can be easily controlled without decrypting the encrypted images.

As shown in **Figs. 6** and **7**, it is difficult to recognize the original image and the dictionary from the encrypted ones, and it would be computationally expensive to obtain the original image and dictionary from the encrypted ones without knowledge of the private key. The authorized user can recover the image $\dot{Y} = Q_p^* \hat{D} \hat{X}$ based on the encrypted dictionary $\hat{D}$ and its sparse representation $\hat{x}$, as shown in **Fig. 8(a)**. The image cannot be recovered by an unauthorized user, as shown in **Fig. 8(b)**.

Moreover, the trained dictionary also demonstrates

*1 JPEG: JPEG is a standard format for image compression developed by the Joint Photographic Experts Group.

Fig. 6. Image before (left) and after (right) encryption.

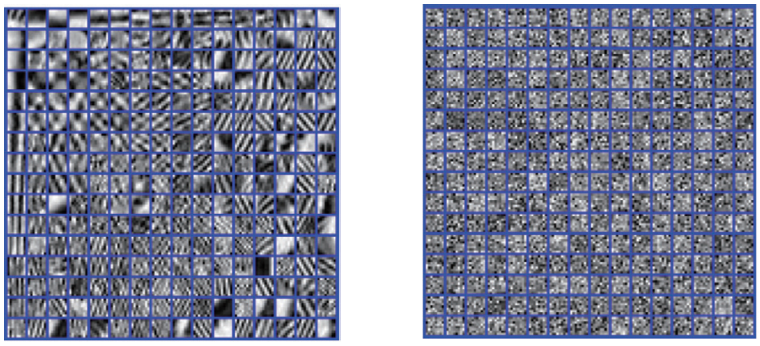


Fig. 7. Dictionary before (left) and after (right) encryption.

(a) Authorized user (PSNR=39.28) (b) Unauthorized user (PSNR=10.40)

PSNR: peak signal-to-noise ratio

Fig. 8. Recovered images.

a representative capability. The rate-distortion performance (average sparsity ratio vs. decoded/decrypted image quality peak signal-to-noise ratio (dB)) when compared with overcomplete discrete cosine transform

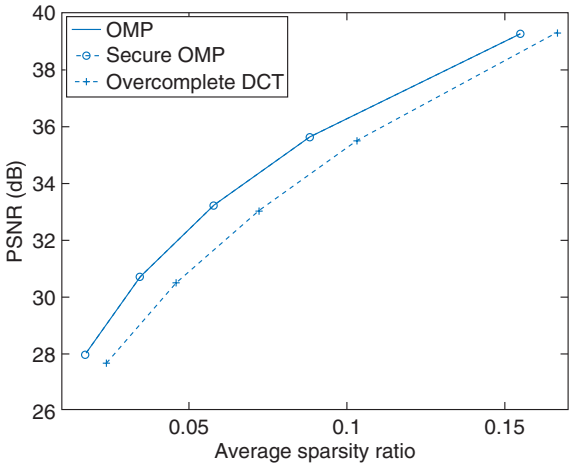


Fig. 9. Rate-distortion performance.

(DCT) is plotted in **Fig. 9**. The average sparsity ratio is defined as the ratio of the number of sparse representations to the number of atoms in the dictionary. It can be seen that the proposed secure sparse coding (secure OMP) can represent the image with fewer atoms than overcomplete DCT. Furthermore, it is confirmed that the proposed secure sparse coding yields the same results as the unencrypted version of sparse coding (OMP) [9].

4. Application to face recognition

Face recognition has been a prominent biometric technique for identity authentication in a wide range of areas and applications, for example, public security and virtual reality. While the integration of face recognition and the edge/cloud network generates an extra degree of freedom for performance enhancement, significant concerns have been raised about privacy, as such biometric information could be misused without permission.

Our objective is to construct a secured framework to reduce the computation demands at each device, while taking advantage of this benefit to produce a more accurate face recognition result. To this end, we preserve privacy by deploying secure sparse coding, which enables dictionaries/recognition results to be trained/drawn from the encrypted images. We further prove both theoretically and through simulation that such encryption will not affect the accuracy of face recognition. To fully utilize the multi-device diversity, we extract deeper features in an intermediate space, which is expanded according to the dictionaries from each device, and perform classification in this new feature space to combat noise and modeling errors. This approach is demonstrated to achieve higher correctness of predictability through simulation results.

In the ensemble training process, as shown in **Fig. 10**, we jointly train a discriminative dictionary and classifier parameter based on the encrypted data at each edge server. Then we extract the decision templates for each class of individuals to be recognized at the remote cloud in order to efficiently combat noise and modeling errors.

In the recognition process, as shown in **Fig. 11**, we devise a pairwise similarity measurement, based on which we compare the current decision profile for a testing sample with each of the formulated decision templates. The closest match will produce the classification result.

We investigate the performance of the proposed framework by simulation. The performance improvement achieved by exploiting the multi-device diversity through ensemble learning is shown in **Fig. 12**. The performance improvement is significant when the number of devices is large due to the extra degree of freedom. In addition, we verified that by adopting random unitary transform, the result of face recognition is not affected, which proves that the proposed framework operates on a secured plane without any performance degradation.

We also show in Fig. 12 a performance comparison with a deep learning based algorithm, in which a 5-layer convolutional neural network was adopted and principal component analysis (PCA) was deployed to learn filter kernels in order to extract more discriminative features. Even though the deep learning based algorithm outperforms the proposed framework by 0.7% in terms of recognition accuracy, it requires 67% more dictionary training samples for each class. Significantly, when there are 30 dictionary training samples for each class, which is the most common setting when evaluating the performance on the dataset we used, the proposed framework dominates by over 4%. When there are only 10 training samples per class, which is reasonable in real-world settings due to the scarcity of fine-grained and manually labeled data, the proposed framework dominates by over 12%.

The performance comparison in terms of computational complexity is indicated in **Table 1**. Even though the deep learning based algorithm adopts: 1) PCA instead of a stochastic gradient descent to learn filter kernels and 2) a hashing method to simplify the

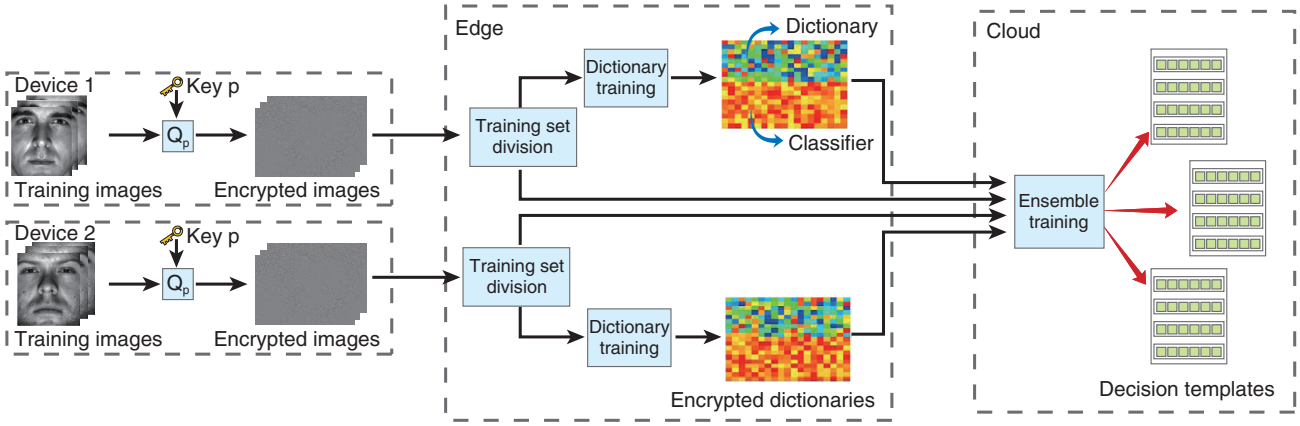


Fig. 10. Ensemble training.

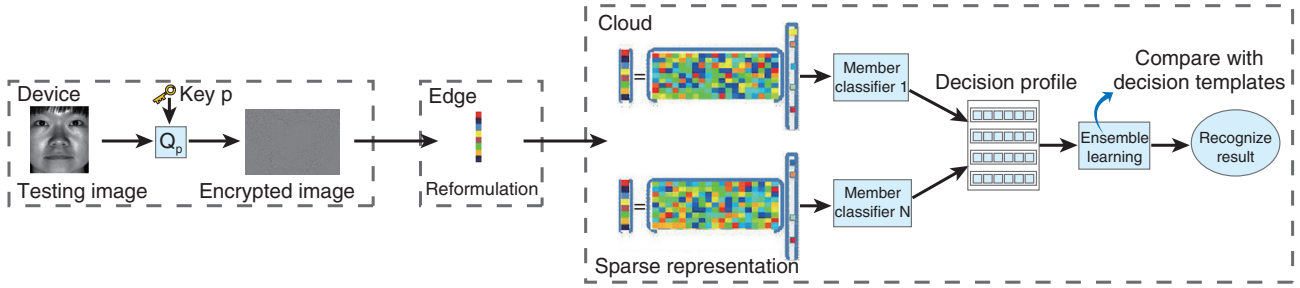


Fig. 11. Recognition process.

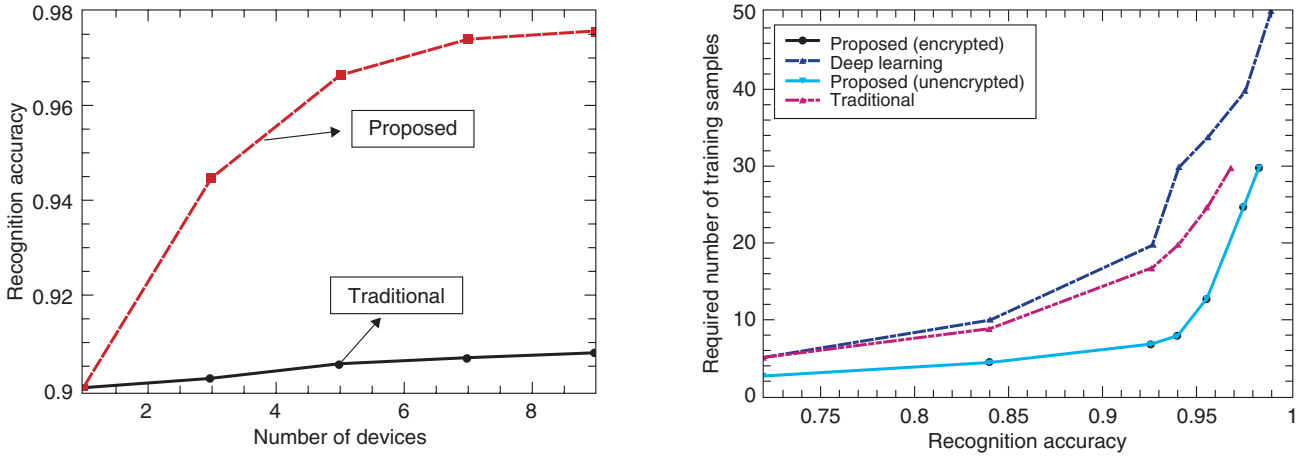


Fig. 12. Multi-device diversity and performance comparison.

nonlinear processing layer in order to reduce the computational complexity, it still requires a very long training time under such a small database. The proposed algorithm is extremely fast in terms of testing

Table 1. Comparison of running time.

	Training time (s)	Testing time (s)
Proposed	7.29	1.64×10^{-3}
Deep learning	5780	1.20
Traditional	4.84	1×10^{-4}

time, which makes it possible to support real-time face recognition applications.

5. Conclusion and future work

We have conducted basic research in the interdiscipline of sparse coding and networking from a security perspective. Specifically, we propose a secure sparse coding scheme with low complexity and demonstrate its application to image compression and face recognition for both preserving privacy and enhancing performance. We plan to further investigate the integration of sparse coding and networking in areas such as online traffic prediction and anomaly detection in order to extend our scientific contribution. Moreover, we are looking for opportunities for practical implementation to support secured real-time multimedia processing related applications, for example, secured face recognition in surveillance cameras, in order to demonstrate its commercial value as well as practical significance.

References

[1] Y. Goyal, A. Mohapatra, D. Parikh, and D. Batra, "Towards Transparent AI Systems: Interpreting Visual Question Answering Models," arXiv:1608.08974, 2016.

[2] B. A. Olshausen and D. J. Field, "Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images," Nature, Vol. 381, pp. 607–609, 1996.

[3] Y. Xu, Z. Li, J. Yang, and D. Zhang, "A Survey of Dictionary Learning Algorithms for Face Recognition," IEEE Access, Vol. 5, pp. 8502–8514, 2017.

[4] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed Deep Neural Networks over the Cloud, the Edge and End Devices," Proc. of the 37th IEEE International Conference on Distributed Computing Systems (ICDCS 2017), pp. 328–339, Atlanta, GA, USA, June 2017.

[5] Y. Mao, J. Zhang, S. Song, and K. Letaief, "Stochastic Joint Radio and Computational Resource Management for Multi-user Mobile-edge Computing Systems," IEEE Trans. Wireless Commun., Vol. 16, No. 9, pp. 5994–6009, 2017.

[6] Y. Saito, I. Nakamura, S. Shiota, and H. Kiya, "An Efficient Random Unitary Matrix for Biometric Template Protection," Proc. of Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS 2016) and 17th International Symposium on Advanced Intelligent Systems (ISIS 2016), pp. 366–370, Sapporo, Japan, Aug. 2016.

[7] T. Nakachi, Y. Bandoh, and H. Kiya, "Secure Dictionary Learning for Sparse Representation," The 27th European Signal Processing Conference (EUSIPCO 2019), Coruña, Spain, Sept. 2019, accepted.

[8] J. Zepeda, C. Guillemot, and E. Kijak, "Image Compression Using Sparse Representations and the Iteration-tuned and Aligned Dictionary," IEEE J. Sel. Topics Signal Process., Vol. 5, No. 5, pp. 1061–1073, 2011.

[9] T. Nakachi and H. Kiya, "Practical Secure OMP Computation and Its Application to Image Modeling," Proc. of the 2018 International Conference on Information Hiding and Image Processing (IHIP 2018), pp. 25–29, Manchester, UK, Sept. 2018.

Yitu Wang

Research Engineer, NTT Network Innovation Laboratories.

He received a B.S. and Ph.D. from Zhejiang University, Hangzhou, China, in 2013 and 2018. From August to November 2014, he was a visiting scholar with the University of Paris-Sud, Orsay, France. He joined NTT Network Innovation Laboratories in 2019. His research interests are mainly focused on communication networks and statistical data processing.

Takayuki Nakachi

Senior Research Engineer, Supervisor, NTT Network Innovation Laboratories.

He received a Ph.D. in electrical engineering from Keio University, Tokyo, in 1997. Since joining NTT, he has been researching super-high-definition image/video coding and media transport technologies. In 2006–2007, he was a visiting scientist at Stanford University, USA. His current research interests include communication science, information theory, and signal processing. He received the 26th TELECOM System Technology award in 2010, the 6th Paper Award of the Journal of Signal Processing in 2012, and the Best Paper Award of the Institute of Electrical and Electronics Engineers (IEEE) International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) 2015. He is a member of IEEE and the Institute of Electronics, Information and Communication Engineers (IEICE).

Report on the 22nd Global Standards Collaboration (GSC-22) Meeting

Hideyuki Iwata

Abstract

The 22nd meeting of the Global Standards Collaboration (GSC-22) was held March 26–27, 2019, in Montreux, Switzerland. The purpose of the GSC is to enable standards developing organizations (SDOs) to share information, avoid duplication of work, and promote standardization activities. Eighty-five people from 12 SDOs participated in the meeting and discussed two strategic topics: connected citizens & smart sustainable cities, and artificial intelligence.

Keywords: Global Standards Collaboration, smart city, artificial intelligence

1. Meeting overview

The 22nd meeting of the Global Standards Collaboration (GSC-22), held March 26–27, 2019, in Montreux, Switzerland, was attended by 85 delegates from 12 standards developing organizations (SDOs): ARIB (Association of Radio Industries and Businesses) of Japan, ATIS (Alliance for Telecommunications Industry Solutions) of the USA, CCSA (China Communications Standards Association), European Telecommunications Standards Institute (ETSI), IEC (International Electrotechnical Commission), the Institute of Electrical and Electronics Engineers (IEEE) Standards Association, ISO (International Organization for Standardization), ITU (International Telecommunication Union), Telecommunications Industry Association (TIA) of the USA, TSDSI (Telecommunications Standards Development Society, India), TTA (Telecommunications Technology Association) of Korea, and The Telecommunication Technology Committee (TTC) of Japan. Delegates from individual SDOs reported on their latest activities and high-priority issues and discussed two strategic topics: connected citizens & smart sustainable cities, and artificial intelligence (AI).

2. Activity reports by individual SDOs

The 12 SDOs reported on their latest activities and high-priority issues.

It was found that a number of SDOs are working on several topics in duplication, such as smart cities, the Internet of Things (IoT), AI, security, and fifth-generation mobile communications (5G). The meeting participants came to a common understanding that it is important for standardization activities to respond to market trends in a timely manner and that SDOs need to collaborate in order to minimize duplicated work.

3. Main themes of GSC-22

This meeting singled out two themes for discussion: smart sustainable cities and AI. Representatives of SDOs made presentations and participated in panel discussions on these themes.

3.1 Smart sustainable cities

A latent problem that stands in the way of realizing a smart city is the absence of a common mechanism by which cities exchange data. The SDOs agreed on the need for continued discussions on the development of guidelines and standards that enable seamless data exchange and mutual operation.

Prominent among the issues reported by SDOs is a lack of interoperability (due to a silo mentality) among different IoT/smart city platforms. Therefore, the participants found it urgent to standardize the

platforms and data APIs (application programming interfaces).

3.2 AI

The participants shared information about the needs of different regions and about sustained activities related to the application of AI and machine learning in the fields of 5G, healthcare, and industrial manufacturing. Some issues addressed at the meeting were not only the technical aspects of AI but also latent problems related to security, privacy, reliability, ethics, social concerns, and regulations. The need for GSC members to cooperate to tackle the ethical and social aspects of information and communication technology systems, services, and technologies was recognized. IEEE reported that it is driving a global project on the ethics of autonomous & intelligent systems, and this project is intended to make designs ethically acceptable. It has formulated the IEEE P7000 series as standards that comprehensively cover ethically conscious AI and autonomous systems.

ETSI technical groups are studying various AI applications. For example, the Industry Specification Group is studying Experiential Networked Intelligence, which uses AI for data collection and analysis, Zero Touch Network and Service Management, which applies AI to automation and operation of networks, and Multi-access Edge Computing, which releases network edges and provides network and context information. The Technical Committee is studying Intelligent Transport Systems, which applies AI to autonomous vehicles.

The chairman of the Security Working Group of TTC (Japan) gave a presentation on transparency and trustworthiness in AI and machine learning. As AI technology spreads and AI-based decision-making penetrates a broad spectrum of activities, it has been pointed out that the use of improper or biased data can lead to inappropriate decision-making. In addition, deliberate maneuverings of AI engines can lead to decisions favorable solely to a specific group of people. He also pointed out that it is necessary to heed the society-oriented basic principles that the Japanese government has defined for human-centric AI: human-centric, education, privacy, security, fair competition, fairness, accountability, transparency, and innovation. The compilation of guidelines for the use of AI systems and the provision of a reporting function in AI systems are essential parts of the effort to promote AI security standardization. He also identified the needs for the internal assessment of training data, for the examination of output data obtained from standardized training data, for the detection of unbalanced output data, and for technical specifications for the trustworthy framework for the use of AI.

4. Next meeting

The next meeting (23rd meeting) will be hosted by TIA (USA) and held in 2020.

Hideyuki Iwata
General Manager, Standardization Strategy, Research and Development Planning Department, NTT.

He received a Ph.D. in electrical engineering from Yamagata University in 2011. From 1993 to 2000, he conducted research on high-density and aerial optical fiber cables at NTT Access Network Service Systems Laboratories. Since 2000, he has been responsible for standardization strategy planning for NTT research and development. He has been a delegate of IEC Subcommittee 86A (optical fiber and cable) since 1998 and of the ITU-T (Telecommunication Standardization Sector) Telecommunication Standardization Advisory Group (TSAG) since 2003. He is a vice-chair of Working Group Policy and Strategic Coordination (WG PSC) and the Expert Group on Bridging the Standardization Gap (EG BSG) in the Asia-Pacific Telecommunity Standardization Program Forum (ASTAP). He received an award from the IEC Activities Promotion Committee of Japan in 2004, the ITU Association of Japan (ITU-AJ) International Activity Encouragement Award in 2005, ITU-AJ International Cooperation Award in 2012, an award for contributions to an information and communication technology (ICT) development project at the Asia-Pacific Telecommunity ICT Ministerial Meeting in 2014, the ITU-AJ Accomplishment Award in 2018, and the TTC Chairman's Prize in 2019.

External Awards

IEICE Communications Society Excellent Paper Award
Winner: Yoshitaka Enomoto, Tetsuya Iwado, Takashi Goto, Masaki Waki, Toshio Kurashima, and Yoshiyuki Kajihara, NTT Access Network Service Systems Laboratories
Date: May 13, 2019
Organization: The Institute of Electronics, Information and Communication Engineers (IEICE)

For "Design and Performance of Aerial Line Structure Inspection Support System with Mobile Mapping System."
Published as: Y. Enomoto, T. Iwado, T. Goto, M. Waki, T. Kurashima, and Y. Kajihara, "Design and Performance of Aerial Line Structure Inspection Support System with Mobile Mapping System," IEICE Trans. Commun. (Japanese Edition), Vol. J100-B, No. 12, pp. 995–1003, 2017.

Laser Society of Japan Achievement Award (Paper Prize)
Winner: Atsushi Ishizawa, Tadashi Nishikawa, Kenichi Hitachi, and Hideki Gotoh, NTT Basic Research Laboratories
Date: May 31, 2019
Organization: The Laser Society of Japan

For "Ultra-precise Frequency Conversion Using an Electro-optic-modulation Frequency Comb."
Published as: A. Ishizawa, T. Nishikawa, K. Hitachi, and H. Gotoh, "Ultra-precise Frequency Conversion Using an Electro-optic-modulation Frequency Comb," The Review of Laser Engineering, Vol. 46, No. 2, pp. 80–85, 2018.

AVM Award
Winner: Masaaki Matsumura, NTT Media Intelligence Laboratories
Date: June 14, 2019
Organization: The Special Interest Group of Audio Visual and Multimedia information processing, Information Processing Society of Japan (IPSJ-AVM)

For "A Study for Prediction of Heating and Strain Using Audience Behavior."
Published as: M. Matsumura, A. Kameda, M. Isogai, H. Noto, and H. Kimata, "A Study for Prediction of Heating and Strain Using Audience Behavior," IPSJ SIG Technical Report, Vol. 2018-AVM-101, No. 10, 2018.

AVM Award
Winner: Shoichiro Takeda, NTT Media Intelligence Laboratories
Date: June 14, 2019
Organization: IPSJ-AVM

For "A Study of Quality of Experience Assessment for Video Magnification."
Published as: S. Takeda, A. Kameda, M. Isogai, and H. Kimata, "A Study of Quality of Experience Assessment for Video Magnification," IPSJ SIG Technical Report, Vol. 2019-AVM-104, No. 9, 2019.

Young Researcher Award
Winner: Takashi Hosono, NTT Media Intelligence Laboratories
Date: June 28, 2019
Organization: The Institute of Image Electronics Engineers of Japan (IIEEJ)

For "Depth Edge Based Objectness Metric for Generating Instance Candidate Regions."
Published as: T. Hosono, S. Tarashima, J. Shimamura, and T. Kinebuchi, "Depth Edge Based Objectness Metric for Generating Instance Candidate Regions," Proc. of Visual/Media Computing Conference 2018, Yamagata, Japan, June 2018.

Outstanding Research Presentation Award
Winner: Shoichiro Takeda, NTT Media Intelligence Laboratories
Date: June 29, 2019
Organization: The Special Interest Group of Computer Graphics and Visual Informatics, IPSJ (IPSJ-CG)

For "Local Riesz Pyramids for Faster Phase-based Video Magnification."
Published as: S. Takeda, M. Isogai, S. Shimizu, and H. Kimata, "Local Riesz Pyramids for Faster Phase-based Video Magnification," IPSJ SIG Technical Report, Vol. 2019-CG-173, No. 4, 2019.

Young Engineer Excellent Presentation Award
Winner: Takashi Miwa, NTT Device Innovation Center
Date: July 5, 2019
Organization: Japan Association of Corrosion Control

For "New Two-layer Paint System with Zinc Rich Paint."
Published as: T. Miwa, A. Ishii, and H. Koizumi, "New Two-layer Paint System with Zinc Rich Paint," Proc. of the 39th Corrosion Control Conference, pp. 75–80, Tokyo, Japan, July 2018.

Young Engineer Excellent Presentation Award
Winner: Azusa Ishii, NTT Device Innovation Center
Date: July 5, 2019
Organization: Japan Association of Corrosion Control

For "Effects of Water Spray on the Degradation Behavior of Poly(ethylene terephthalate) under Accelerated Weathering Test."
Published as: A. Ishii, T. Miwa, M. Watanabe, and S. Oka, "Effects of Water Spray on the Degradation Behavior of Poly(ethylene terephthalate) under Accelerated Weathering Test," Proc. of the 39th Corrosion Control Conference, pp. 23–28, Tokyo, Japan, July 2018.

Papers Published in Technical Journals and Conference Proceedings

Experimental Demonstration of Secure Quantum Remote Sensing

P. Yin, Y. Takeuchi, W. Zhang, Z. Yin, Y. Matsuzaki, X. Peng, X. Xu, J. Xu, J. Tang, Z. Zhou, G. Chen, C. Li, and G. Guo

arXiv:1907.06480 [quant-ph], July 2019.

Quantum metrology aims to enhance the precision of various measurement tasks by taking advantage of quantum properties. In many scenarios, precision is not the sole target; the acquired information must be protected once it is generated in the sensing process. Considering a remote sensing scenario where a local site performs cooperative sensing with a remote site to collect private information at the remote site, the loss of sensing data inevitably causes private information to be revealed. Quantum key distribution is known to be a reliable solution for secure data transmission; however, it fails if an eavesdropper accesses the sensing data generated at a remote site. In this study, we demonstrate that by sharing entanglement between local and remote sites, secure quantum remote sensing can be realized, and the secure level is characterized by asymmetric Fisher information gain. Concretely, only the local site can acquire the estimated parameter accurately with Fisher information approaching 1. In contrast, the accessible Fisher information for an eavesdropper is nearly zero even if he/she obtains the raw sensing data at the remote site. This achievement is primarily due to the nonlocal calibration and steering of the probe state at the remote site. Our results explore one significant advantage of "quantumness" and extend the notion of quantum metrology to the security realm.

Proposal and Verification of Auto Calibration Technique for Bias Control Circuit Connecting to Imperfect IQ-modulator

H. Kawakami, S. Kuwahara, and Y. Kisaka

Proc. of the 24th OptoElectronics and Communications Conference (OECC 2019), ThC2-2, Fukuoka, Japan, July 2019.

We show that imperfection in an IQ-modulator degrades the accuracy of the auto bias control circuit connected to the modulator's complementary port, and propose an auto calibration technique that can effectively suppress this degradation.
