NTT DOCOMO
Technical Journal
Technical Journal
Vol.23 No.4 | Apr. 2022

DOCOMO Today

- Becoming a Green Carrier

Technology Reports (Special Articles)

Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

- Natural Language Processing for Realizing DX
- Multifunctional Automatic News Article Summarization AI System for Efficient Summarization
- A Recommendation Engine Using Time Series Prediction Models of User Behavior
- Providing Image Recognition AI via the DOCOMO Image Recognition Platform
- "Generic POI Recommendation": A Brand-new Deep Learning Approach for Discovering Potential Sightseeing Spots
- Self-driving Support System toward Personal Mobility Using Edge AI-compatible 5G Device

Technology Reports

- SR-based Routers in 5G MBH
- Migration to ETSI NFV Stage 3 Specification-compliant Multivendor MANO Configuration on Network Virtualization Platform

Standardization

- Standardization of Frameworks for Industrial Application Enablement in 3GPP

NTT docomo

Becoming a Green Carrier

Mobile telecommunications carriers in Japan currently emit an amount of CO_2 equivalent to that of approximately 10 million private vehicles/year*. The amount of these emissions, which originate in the charging of mobile terminals as well as the operation of base stations and the manufacture and delivery of terminals, is extremely large, and the reduction of CO_2 emissions is a great social responsibility for all of us. Movements toward achieving a decarbonized society are accelerating throughout the world, and NTT DOCOMO, which views its response to climate change as an important corporate issue, declared its commitment to carbon neutrality by 2030 in September 2021 [1].

This declaration states that NTT DOCOMO will undertake this challenge within the company, throughout its value chain, and together with its customers and partner companies. First, we will achieve carbon neutrality within the company in a number of ways, such as by developing technologies and implementing facilities for reducing power consumption in the network and by procuring renewable energy from NTT DOCOMO's own solar power plants. Next, in the value chain, we will promote the use of renewable energy at docomo Shops by installing solar panels at their premises. Finally, for customers and partner companies, we will roll out new services designed to contribute to decarbonization such as "Green 5G" that is essentially powered by renewable energy and the sale of "green power" via the "docomo Denki" electricity service. We will also contribute to carbon neutrality throughout society by providing environmentally conscious products and services from partner companies for use by customers in conjunction with NTT DOCOMO services. Through these initiatives, NTT DOCOMO aims to become a green carrier.

In line with this declaration, NTT DOCOMO R&D will strive to develop technologies for reducing CO_2 emissions originating in the network and services. For the network, there are a wide variety of technologies that can contribute to a reduction in CO_2 emissions. These include enhanced base-station sleep functions and reduced power consumption in 5th Generation mobile communications system (5G) equipment as well as NTT's Innovative Optical and Wireless Network (IOWN) that includes a next-generation network and information-processing infrastructure. As for services, we plan to contribute to advanced solutions in various fields by developing service-related technologies that can encourage people to undergo a behavioral transformation toward a low-carbon lifestyle. We can expect the amount of CO_2 emissions to increase in step with an increase in the volume of communications. With this being the case, achieving a decarbonized society will require even more innovative developments in the network and services, so I feel that even more will be expected of NTT DOCOMO R&D.

In addition to the above, R&D General Affairs Department that I belong to is working on creating a safe-and-secure R&D environment at the DOCOMO R&D Center located in Yokosuka Research Park. In face of the COVID-19 pandemic, we are enhancing diverse measures to prevent the spread of infections while providing workplace vaccinations. A major theme of late is how best to achieve decarbonization at the DOCOMO R&D Center, and measures for reducing power consumption are being phased-in such as installing air-conditioning facilities with high energy-saving performance. We also plan to expand the use of renewable energy by installing large-scale solar panels and turning purchased power effectively into renewable energy through the purchase of non-fossil fuel certificates. Through efforts such as these, I would like to achieve "carbon neutrality from the R&D stage" and help NTT DOCOMO to become a green carrier.

General Manager of
R&D General Affairs Department
Yoshifumi Kurokawa

In closing, I would like to introduce you to the WHARF exhibition hall at the DOCOMO R&D Center. Open to the general public, WHARF visitors can experience for themselves mobile communications of the future and the Smart Life. About 140,000 people, a great number indeed, have so far come to visit WHARF. In addition to elemental technologies making up 5G and the 6th-Generation mobile communications system (6G), WHARF brings together a number of examples of how 5G can be used in a variety of fields such as humanoid robots, eXtended Reality (XR), and esports so that our customers can experience the latest technologies. In the future, we plan to supplement these exhibits with examples of innovations related to our pursuit of carbon neutrality. At the WHARF reception desk, an autonomous-running robot greets and guides visitors, and in this way, I would like to make positive use of the DOCOMO R&D Center as a place for holding trials of new technologies.

REFERENCE

[1] NTT DOCOMO Press Release: "NTT DOCOMO Commits to Carbon Neutrality by 2030, Aiming to Achieve Effectively Zero Greenhouse Gas Emissions—Initiatives Begin on Helping Society as a Whole Achieve Carbon Neutrality Together with its Customers and Partner Companies—," Sep. 2021.

* The equivalent number of vehicles is calculated by summing the amount of CO_2 emissions (FY2019) publically announced by Japanese mobile telecommunications carriers (DOCOMO, KDDI, Softbank, and Rakuten) and dividing by the amount of CO_2 emissions per private vehicle.

NTT DOCOMO
Technical Journal
Vol.23 No.4 Apr. 2022

[Contents]

DOCOMO Today

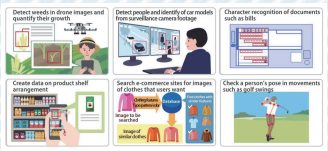
Becoming a Green Carrier Yoshifumi Kurokawa 1

Technology Reports (Special Articles)

Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

Natural Language Processing for Realizing DX 5

DX | Natural Language Processing | Document Classification



(P.36)

Multifunctional Automatic News Article Summarization AI System for Efficient Summarization 15

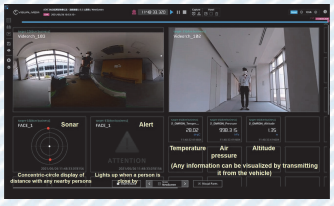
AI | Automatic Summarization | Natural Language Processing Techniques

A Recommendation Engine Using Time Series Prediction Models of User Behavior 28

Behavior Prediction | Recommendations | RNN

Providing Image Recognition AI via the DOCOMO Image Recognition Platform 36

AI | Image Recognition | Deep Learning



(P.54)

“Generic POI Recommendation”: A Brand-new Deep Learning Approach for Discovering Potential Sightseeing Spots 46

Tourist Spot Recommendation System | Tourism Informatics | Neural Network

Self-driving Support System toward Personal Mobility Using Edge AI-compatible 5G Device 54

5G | Edge AI | Self-driving

Technology Reports

(P.74)

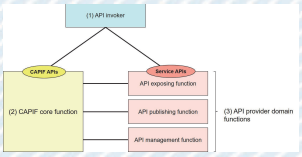
SR-based Routers in 5G MBH 63

MBH | SR-MPLS | TI-LFA

Migration to ETSI NFV Stage 3 Specification-compliant Multivendor MANO Configuration on Network Virtualization Platform 74

Network Virtualization | Orchestration | Upgrade without Service Interruption

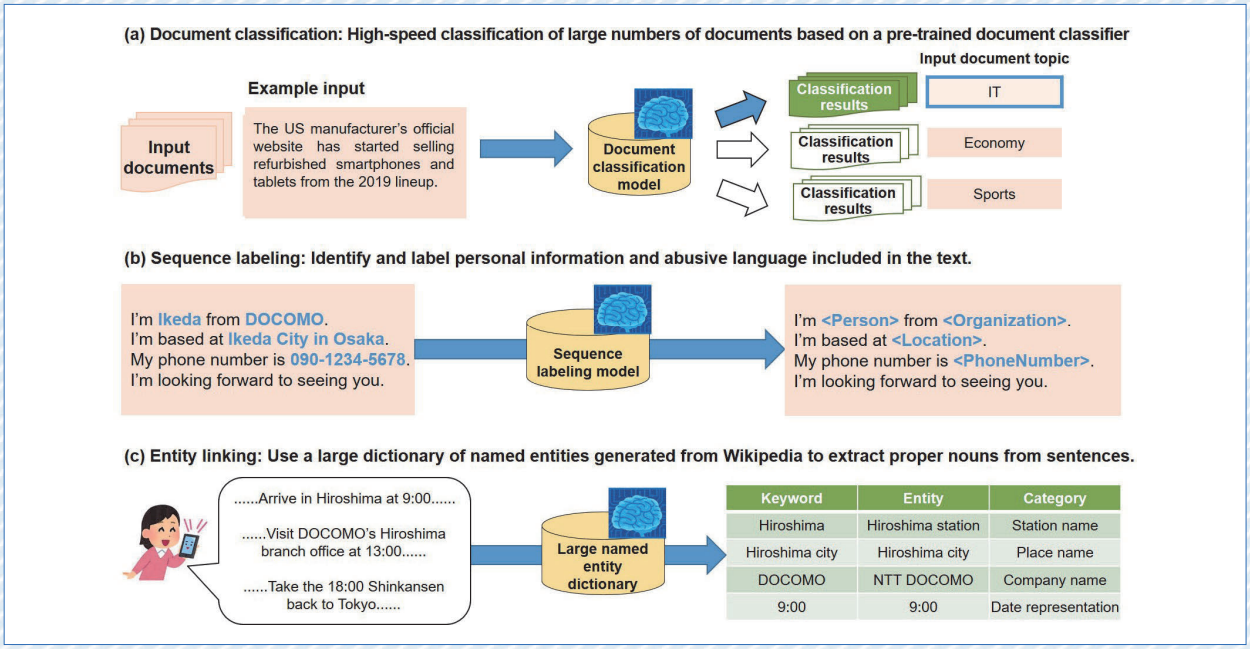
Standardization



(P.87)

Standardization of Frameworks for Industrial Application Enablement in 3GPP 87

CAPIF | SEAL | API



Technology Reports (Special Articles) Natural Language Processing for Realizing DX (P.5)
List of natural language processing AI features

Special Articles on AI

Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications

Today, smartphones have become an integral part of daily life and the launch of 5G services has made a high-speed, low-latency mobile communications environment a reality. Against this background, the research and development of AI has accelerated throughout the world with the aim of making our life more convenient through the use of smartphones and communication services. The NTT DOCOMO Group as well is energetically exploring the use of AI technologies.

The basis of AI is machine learning and statistical analysis. While the fundamental task of AI is to classify data and make suggestions, it has found widespread application.

The NTT DOCOMO Technical Journal periodically features special articles on AI, and in this issue, we introduce activities related to language processing and image recognition in relation to data classification and the research and development of a recommendation engine and self-driving vehicles in relation to behavior support.

These special articles describe a variety of challenges that go beyond the simple application of technology while presenting many application examples that reflect new possibilities in AI. We hope they deepen the reader's understanding of AI and lead to the creation of new forms of AI that go beyond existing frameworks.

(Editorial Office)

Technology Reports (Special Articles)
DX | Natural Language Processing | Document Classification
Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

Natural Language Processing for Realizing DX

Service Innovation Department Yutaro Shiramizu Keisuke Oka Shusuke Tatsumi

In recent years, there has been a great deal of interest in the development of technologies that use AI to help promote DX. In particular, attention has been focused on the field of natural language processing, which is an AI technology that can quickly and accurately process large amounts of text data accumulated by businesses. This article introduces the natural language processing AI being developed by NTT DOCOMO, its technical features, and the GUI tools that support the use of AI. It also discusses actual implementations of this technology, its synergistic effects with RPA tools, and its future prospects.

1. Introduction

In recent years, there has been a government-led push to reform the way people work, and there is a strong demand for the use of digital technologies such as Robotic Process Automation (RPA) tools*1 and chatbots*2 to speed up business operations, improve productivity, and reduce workloads. In particular, RPA tools can make significant improvements to the processing of large quantities of text data, including questionnaires and user inquiries, and are becoming increasingly popular, especially in front and back offices. However, with RPA tools alone, it is difficult to achieve the Digital Transformation (DX)*3 of work requiring intelligent judgments, such as text categorization and

*1 RPA tool: A software-based robot that records PC operations as scenarios and automates PC operations.

tagging. There is consequently also an increasing demand for AI-based automation in the field of natural language processing, which can replace human linguistic comprehension.

However, AI has to be diverse because the available computational resources and required performance differ from one project to the next in practical situations. People also have reservations about the security aspects of using AI to process a company's private internal text data in public external Software as a Service (SaaS)*4 applications. Furthermore, departments without system engineers can find it very burdensome to implement the introduction and operation of AI by themselves. Due to issues such as these, the introduction of AI is not always an easy task.

To address these issues, NTT DOCOMO has developed a high-performance natural language processing algorithm that makes full use of the computational resources, and a lightweight and high-speed algorithm that can run on low-spec PCs. These algorithms have made it possible to select and use flexible natural language processing AI technologies that meet the needs of users in diverse scenarios, including cloud and local processing environments, PCs and servers with diverse specifications, and various user security policies. To reduce the burden on operators, we have also developed a Graphical User Interface (GUI)*5 tool that makes these AIs easy to use and fine-tune.

This article describes the development of natural language processing AI and GUI technologies that meet diverse and widely varying needs in practical tasks such as text categorization and the concealment of personal information, as well as in internal and external applications.

2. Natural Language Processing AI to Help Drive DX

2.1 Overview

In recent years, natural language processing has been the subject of academic research all over the world, and many high-performance algorithms have been proposed in studies conducted under well-developed experimental settings. These algorithms are often released as Open Source Software (OSS)*6, so the latest algorithms are readily available for anyone to use. On the other hand, in the business world, there are many restrictions on the availability of computational resources and data, so it is not possible to satisfy the needs of users by applying these algorithms directly. At NTT DOCOMO, we have therefore been working on the development of additional functions and the creation of methods for deploying these functions to meet the diverse needs of real users while using the latest OSS.

2.2 Functions

1) Document Classification

A document classification function is a function that automatically assigns labels to documents (**Figure 1** (a)). NTT DOCOMO provides two types of classifier (lightweight and high-performance) to satisfy users with different needs.

(a) Lightweight classifier

The lightweight classifier meets the needs of local environments using ordinary business PCs for situations such as when the cloud cannot be used due to concerns about

*2 Chatbot: A program that automatically conducts interactive dialog with people via speech or text.
*3 DX: The use of IT technology to revolutionize services and business models, promote business, and change the lives of people for the better in diverse ways.
*4 SaaS: Software that is used remotely via the Internet or other networks.
*5 GUI: An interface that consists of a combination of buttons and icons that are clearly visible and can be operated intuitively.
*6 OSS: Software whose source code is released free of charge for anyone to reuse or modify.

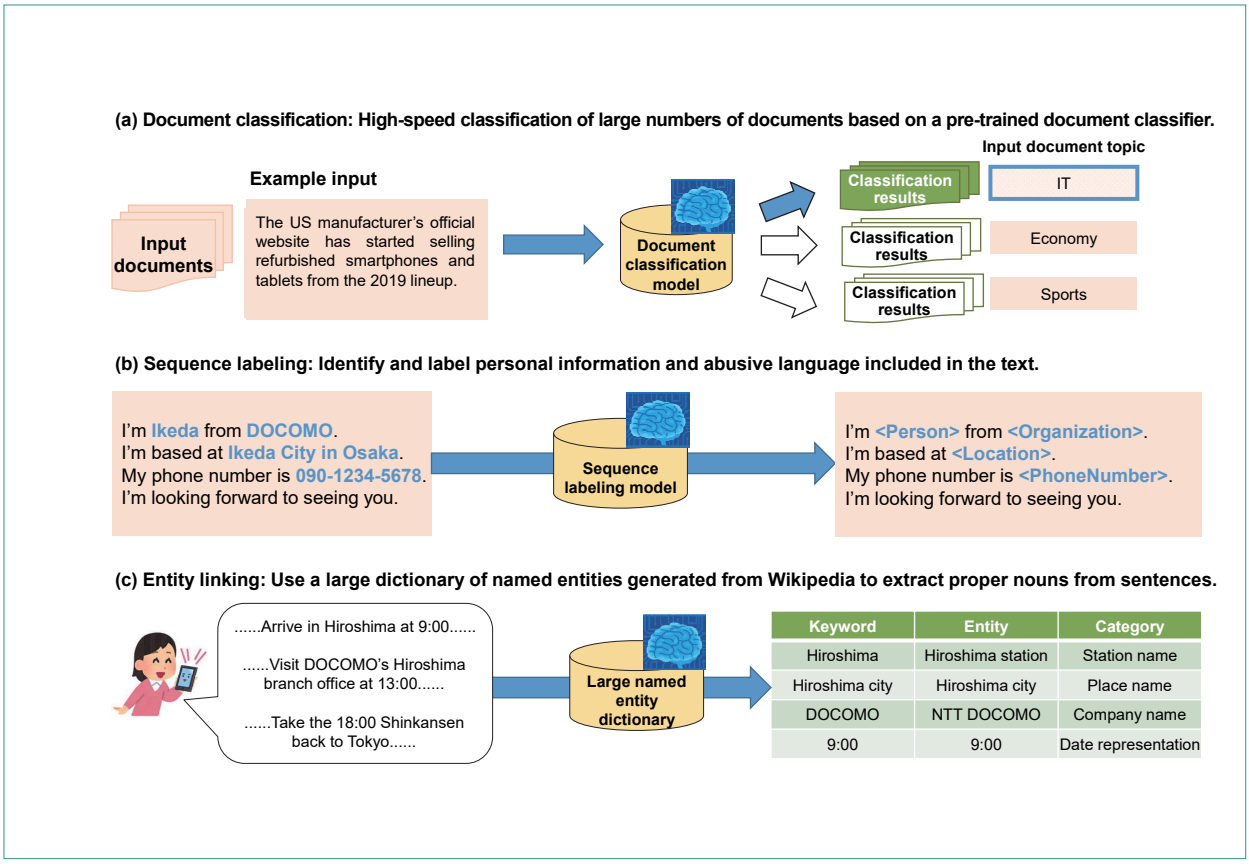


Figure 1 List of natural language processing AI features

cost and security. Its algorithm uses a multilayer perceptron*7 that is simple to implement, computationally inexpensive, and capable of running on a low-spec PC. The features*8 used in this algorithm consist not only of morphemes*9, which are commonly used for this purpose, but also the actual characters used in the text and the appearance of character N-grams*10 within it. This makes it possible to achieve high accuracy even with a small corpus*11, and enables the construction of classifiers that are robust against misspellings and omissions. **Figure 2** shows an example of FAQ classification to illustrate how these features can be used effectively. In FAQ classification, the text of a question entered by a user is automatically classified to match a prepared FAQ item. Here, if the input text contains misspellings, it cannot be classified correctly when using only morphemes. On the other hand, if characters and character N-grams are used, a question can be correctly classified by inferring substrings (like "b, l, u, e, t, o, ..." in the sample shown in the figure) as hints.

(b) High-performance classifier

The high-performance classifier is designed to meet the needs of users who have either

*7 Multilayer perceptron: A type of neural network that implements machine learning algorithms.
*8 Feature: An amount (a numeric value) extracted from data to characterize that data.
*9 Morpheme: The smallest unit of a linguistic expression.
*10 N-gram: Any sequence of n consecutive elements (e.g., words or letters).
*11 Corpus: A language resource consisting of a large volume of text and utterances, etc. collected and stored in a database.

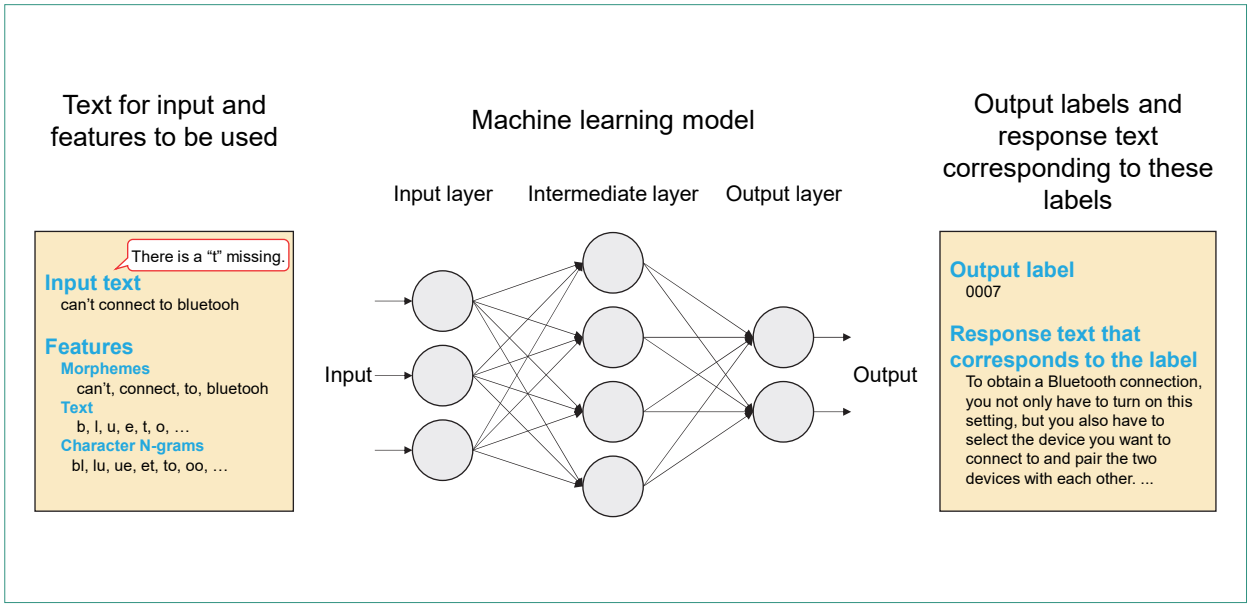


Figure 2 An example where the features of a lightweight classifier work effectively

a cloud server or an on-premises*12 server equipped with a Graphics Processing Unit (GPU)*13, and who want to achieve high performance by making the best use of abundant computational resources. It uses an algorithm called Bidirectional Encoder Representations from Transformers (BERT) [1], which has become the de facto standard for language processing in recent years. BERT requires large amounts of training data and computational resources, but performs very well. Its pre-learning model*14 is based on the NTT version of BERT*15, a technology developed by NTT Human Informatics Laboratories.

NTT DOCOMO has also provided its own support for multi-label classification*16. The NTT version of BERT only supported single-label classification*17, but there is a strong demand for multi-label classification in the business field. To meet this need, NTT DOCOMO altered the label output part of NTT's BERT classifier to provide multi-label output.

2) Sequence Labeling

The purpose of the sequence labeling function is to automatically assign labels to sequential data*18. At NTT DOCOMO, we use this function to mask text including content such as personal information and abusive language (Fig. 1 (b)). In developing this function, we used an open-source natural language processing library called Flair [2], which has a large developer community and can handle lightweight algorithms. In the selection of OSS, the use of lightweight algorithms was an important consideration because users often have to avoid using the cloud for security reasons when processing highly confidential data, and therefore need a solution that can also operate in a local environment. A large developer community is also

*12 On-premises: Refers to an environment where the constituent hardware of a corporate system is owned, operated and maintained by the company itself.

*13 GPU: A processor unit that excels in parallel computing. It is ideal for deep learning processes that requires parallel computation.

*14 Pre-learning model: A model that has been trained unsupervised on a large corpus before supervised learning of the target task.

*15 NTT version of BERT: A BERT model that was pre-trained using data collected by NTT Human Informatics Laboratories.

important, because it makes it more likely that the software will be continuously updated with interface improvements and bug fixes based on feedback from diverse users. We avoided OSS from small developer communities because it tends to have meager documentation and less accumulated know-how, so that the development and operation of systems often depends on an individual developer.

3) Entity Linking

The entity linking function extracts keywords from sentences and then links these keywords to entities. For example, sentences may contain not only common nouns like "desk" and "train", but also proper nouns like "America", "USA", and "United States". Sometimes, different nouns may refer to the same concept (entity) such as the last three examples in the previous sentence, which all refer to the United States of America. Entity linking is a function that mechanically extracts keywords from a sentence (words that convey a unique meaning) and infers the entities to which these keywords refer (Fig. 1 (c)).

(a) Building a dictionary based on data extracted from Wikipedia

At NTT DOCOMO, we perform entity linking by using data extracted from Wikipedia [3]. This data – including each article's body text, anchor text, and number of page views – is statistically processed and used to weight the keywords and construct an entity-keyword dictionary. Using Wikipedia has several advantages: (a) its articles are frequently created and updated, and it responds quickly to new phenomena and popular words, (b) new database dumps*19 are published every day, making the data easy to use, and (c) the cost of maintaining the dictionary is low because the entire process from data extraction to dictionary construction is fully automated.

(b) Category assignment

NTT DOCOMO also assigns categories to entities as higher-level concepts based on an extended named entity hierarchy*20 [4], and classifies the extracted results into about 200 categories before they are used. For example, in a dialogue system, it would not be practical to prepare individual system responses for every single possible user utterance, such as "I want to eat noodles" and "I want to eat an apple." However, if we use the fact that the words "noodles" and "apple" both belong to a "food name" category, it is only necessary to prepare a system response for "I want to eat [food name]," which greatly reduces the burden of creating response scenarios.

(c) Use case

One possible use case of entity linking is the extraction of keywords from news. The use of Wikipedia statistics for entity linking is a good match with keyword extraction from news items, because Wikipedia updates its article content and increases the number of page views to reflect current trends. Furthermore, since it is possible to perform not only simple character string extraction but also to accommodate different ways of representing the same concept, this approach makes it easy to use extracted/linked entities or categories associated with entities as article tags.

*16 Multi-label classification: A classification method where multiple labels can be assigned to a single piece of data.

*17 Single-label classification: A classification method where a single label is assigned to a single piece of data.

*18 Sequential data: Data consisting of a series of elements, such as character strings, audio waveforms, and purchase histories.

*19 Database dump: A file containing the content of a database.

3. GUI Tools to Accelerate the Deployment of Natural Language Processing AI

3.1 Overview

When implementing and using natural language processing AI, it is important to tune the AI to maintain and improve its accuracy. Tuning refers to a series of tasks involved in the construction and continuous updating of AI models, such as verifying the AI accuracy and modifying the training data by performing annotation*21 work. It also involves visual work by operators, which incurs a certain level of human labor costs.

To reduce these labor costs, NTT DOCOMO has also developed GUI tools to support AI tuning. With these tools, it is possible to use simple screen operations to rapidly perform AI training and evaluation, which requires repetitive work such as construction and tuning. In addition, the prediction results of AI trained by users can be used for annotations to adjust the amount of manual effort and prioritize tasks so that work can be performed efficiently. Furthermore, since the constructed AI and data are managed by the system in units of revisions, it is no longer necessary for the user to be aware of complicated information management such as which data was used to construct which AI and when, and how accurate it was. As regards the introduction of these tools, their back-end*22 engine runs on container virtualization technology*23 and thus does not depend on work environment attributes such as a local/cloud server or host OS. These GUI tools are also containerized so that the back-end tools and front-end*24 GUI can both be introduced easily.

In this way, by using GUI tools that are easy to introduce, users can easily build, tune and manage AI systems by performing simple screen operations, which is expected to reduce the hurdles that have to be negotiated in order to set up and use natural language processing AI.

3.2 GUI Tools and Their Functions

The GUI tools provide a suite of functions that are needed for the implementation of AI, including uploading datasets, creating annotation data, learning, and performing evaluations. Everything from the initial construction of AI systems to the maintenance of these systems can be done on-screen via a GUI (**Figure 3**).

The details of the main functions are outlined below.

1) Learning/evaluation

Users can create their own AI systems by using the GUI tools to upload pre-prepared annotated text data to the backend server and perform learning. To judge the accuracy of labels automatically assigned to the input, the user can upload annotated data prepared as correct answers to the GUI tool, whereby the AI can then automatically calculate the correct answer rate and evaluate the accuracy.

In this way, users can easily construct their own AI systems and evaluate their accuracy simply by preparing annotated text data.

2) Tuning

With this function, users can check and correct labels by using trained AI to automatically assign labels to text data. In general, AI can reproduce human judgment by learning from texts labeled by humans. The more texts it has to learn from,

*20 Extended named entity hierarchy: A semantic classification structure wherein words are classified into about 200 categories such as "names of people", "names of cities and towns", and "names of countries". These categories have a structure of up to three levels, such as "place name > name of astronomical body > planet name".

*21 Annotation: Manually annotating data such as text and images

*22 Back-end: The part of a system that runs the GUI. It mainly consists of a processing engine and a system part that connects the engine with the GUI.

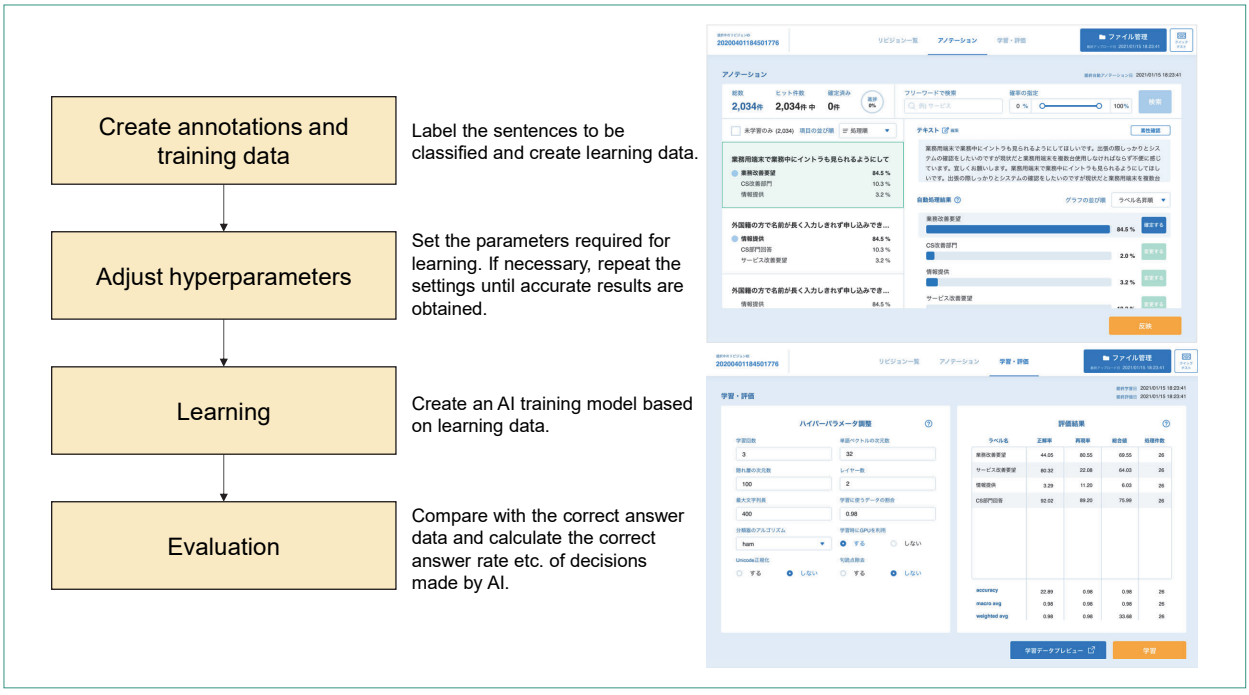


Figure 3 Tuning flow and screenshot of the GUI tools

the better its accuracy.

Once the AI has been trained to some extent, the annotation process can be made more efficient by allowing it to start annotating text, and using human judgment to correct its mistakes. The corrected annotations can then be used to improve the AI accuracy by further training.

As a result, this function makes it possible to improve the accuracy of AI while reducing the cost of creating labeled text data.

It also provides a screen where users can perform actions such as executing learning, evaluating the accuracy of learning, and changing hyperparameters*25 during learning. This makes it easy to adjust parameters*26 during tuning.

3) Revision Control

When tuning an AI system, the complexity of managing previous datasets and evaluation results can often become an operational issue. This tool provides functions whereby trained AI systems and the datasets and evaluation summaries used in learning can be managed in work units called "revisions." In this way, users do not need to be aware of the abovementioned complex management. Instead, the previous datasets and evaluation summaries can be individually managed on the system side, allowing the effects of tuning and differences of data to be analyzed for each dataset.

4. Use cases

4.1 Overview

At NTT DOCOMO, we are not only developing natural language processing AI and GUI tools, but we are also promoting DX using natural language processing AI both inside and outside the

*23 Container virtualization technology: A technology in which an application itself and all its necessary files are packaged together as a "container" and run on a process called a container engine.

*24 Front-end: The part of the system that the user sees and interacts with. This mostly corresponds to the GUI.

*25 Hyperparameter: A value that is set during training. Since the performance varies depending on the setting values, they should be optimized for the best performance.

*26 Parameter adjustment: The selection of relevant setting values for the best performance.

company.

Internally, we are cooperating with the Customer Satisfaction (CS) department, which collects text-based feedback from users and employees. We are working to improve their operations and help them achieve greater user satisfaction through measures such as constructing and supporting the introduction of natural language processing AI, and sharing the accumulated results of manual classification.

Outside the company, we are also working with the corporate sales division to develop RPA-AI collaborative solutions, and we are conducting a wide range of demonstration experiments using not only natural language processing but also Optical Character Recognition (OCR) and speech recognition[*27] technologies.

4.2 Working with the In-house CS Department

NTT DOCOMO collects feedback from users and employees at its shops and call centers. After it has been anonymized so that specific individuals cannot be identified, we use this data to create better services and operations that lead to increased user and employee satisfaction. For example, questionnaires submitted by users are read by specialist staff who compile them into units such as "requests" and "plaudits" before they are shared with related departments. However, a large number of questionnaires are received every day, and it takes a lot of time to read them all properly. Since this process relies on human judgment, it is inevitable that the accuracy will decrease due to oversights or misclassifications. Another issue is that it is difficult to equalize the quality of this work because the classification results are dependent on the knowledge and experience of staff.

NTT DOCOMO has therefore been working with the in-house CS department to promote the automatic classification of questionnaires by introducing natural language processing AI with the aim of improving operational efficiency by promptly and appropriately responding to user feedback. Since the collected questionnaires were used and analyzed within a closed in-house network, it would have been difficult to have them automatically processed by an external high-spec server. We therefore developed the abovementioned lightweight classification algorithm and made it available as AI for use in classifying questionnaire responses. Since data from previous questionnaires had already accumulated by manual classification, the CS department used it as AI learning data.

By introducing natural language processing AI into this work, we were able to automate most of the work that had previously been done manually for a long time. As a result, the questionnaire processing workload was greatly reduced. In the future, in addition to the automatic classification of questionnaires, we also aim to achieve greater user satisfaction by introducing AI that automatically conceals personal information such as names and addresses that users have mistakenly entered.

4.3 Solutions for Linking RPA Tools and AI

In recent years, solutions to automate more complex tasks (Cognitive Automation[*28]) have been explored by combining RPA tools with "brains" implemented using AI for natural language processing or OCR. As part of the introduction and expansion

*27 Speech recognition: Technology that analyzes human speech in order to convert it into text or infer people's emotions.

*28 Cognitive Automation: A technology that combines RPA tools with AI tools such as natural language processing, image recognition and voice recognition to enable the automation of judgment-based work that is performed by humans.

of our WinActor RPA tool and our delivery of RPA-AI collaborative solutions, NTT DOCOMO has also been working with the Nara Prefectural General Medical Center on demonstrations of an experimental system for automatically inputting electronic medical records into existing systems.

Although humans are required to do some of the work in the medical field, such as performing medical examinations and tests, there are also many tasks that can be made more efficient by IT, such as creating medical records for each patient and inputting data into electronic medical record systems. However, especially when inputting data to the system, it is necessary to have some level of understanding of the text data before it can be transcribed to the proper location, so it is not easy to perform this task with RPA tools alone. In addition, since this process involves handling personal information, it is preferably performed within the confines of a local network between work PCs rather than by a SaaS solution that sends the data over an external network. In addition, hospitals often already have databases and systems up and running, and these have to be used as they are without any modification.

To accommodate these requirements, we constructed a lightweight natural language processing AI tool that can run locally and an RPA tool that requires only a PC connected to the hospital network and does not require an external server. We linked these systems together to automate the work of transcribing records into the system that had previously been performed by humans. By doing so, we were able to eliminate excess work by reducing the workload of medical personnel without making major changes to the later stages of the workflow. We also introduced new voice recognition software with the aim of improving the efficiency of data creation by making it possible to convert printed patient records into electronic data simply by reading them out aloud.

In this demonstration experiment, we evaluated not only quantitative effects such as the reduction of the time required to enter data into the electronic medical record system, but also the qualitative effects mentioned by people responsible for entering this data, who said that it reduced the incidence of data entry errors and gave them more time to spend with patients. In this way, the introduction of the RPA-AI collaborative solution made it possible for hospital staff to concentrate on their core business, and produced results leading to the creation of new value for customers. These results strongly suggest that DX is not only able to deliver cost reductions, but can also contribute to improving the quality of work itself.

5. Conclusion

We have discussed the development of natural language processing algorithms for delivering DX and GUI tools for improved operability and convenience, and the deployment and future prospects of these tools both within NTT DOCOMO and in other businesses. By developing a variety of natural language processing AI tools supporting diverse methods and algorithms, we have made it possible for users to select the appropriate AI according to their applications, requirements, and execution environments. At the same time, we sought to make operating AI in the field less burdensome (including making it easier to train systems, evaluate

their performance, and improve their accuracy) by developing GUI tools that incorporate functional requests from the operating department. We have also shown that the natural language processing AI we developed is not only able to quantitatively reduce the workload of operators, but also has a strong affinity with RPA tools and the improvement of work quality due to improvement of the workflow.

In the future, in parallel with the expansion of back-end functions such as the addition of new algorithms, we hope to promote front-end GUI tools for creating use cases inside and outside the company and increasing the number of organizations that use natural language processing AI.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. L. and K. Toutanova: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. of NAACL, May 2019.
[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter and R. Vollgraf: "FLAIR: An easy-to-use framework for state-of-the-art NLP," Proc. of NAACL, Jun. 2019.
[3] Wikipedia home page. https://en.wikipedia.org/wiki/Main_Page
[4] Shinra Project: "Extended named entity." http://ene-project.info/?lang=en

Technology Reports (Special Articles)
AI | Automatic Summarization | Natural Language Processing Techniques
Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

Multifunctional Automatic News Article Summarization AI System for Efficient Summarization

DOCOMO Beijing Communications Laboratories Xuhong Guo Issei Nakamura Anxin Li
Service Innovation Department Hiroshi Fujimoto

Nowadays, many media outlets deliver summaries of news, which helps readers to understand news efficiently. However, manually writing a summary of each news article requires specialized skills, and leads to the issue of securing the human resources. NTT DOCOMO has developed a multifunctional AI system for automatic news article summarization. This system generates summaries based on the user's intention, reducing the time required for summarization and addressing the shortage of human resources.

1. Introduction

Nowadays, many media outlets provide simple summaries of news articles to present the main content to readers in an easy-to-understand manner. However, summaries often have a limit on the number of sentences or characters due to the area constraint on media outlets, and manually condensing news article into summaries with appropriate length requires specialized skills, in addition to handling large amounts of workloads. Therefore, it is necessary to train staff for a certain period of time, which is an issue from the standpoint of securing human resources.

To address this issue, in recent years a variety of automatic summarization systems using AI

technology, which can improve the efficiency of summarization, have been released. Many of these systems use extractive summarization or abstractive summarization techniques. Extractive summarization extracts key sentences from the source text, whereas abstractive summarization creates new sentences by deleting or adding some words or phrases from the source text. Examples of extractive summarization and abstractive summarization are shown in **Figure 1**. The examples in all figures are provided only in Japanese since our system supports only Japanese at present.

Although having potential to generate high-quality news summaries in a similar way as a human, abstractive summarization has the following issues. Since some words or phrases are deleted or added to meet the required length of the summary, conventional technology often generates the summary with grammar errors. The summary may also not satisfy length requirements even if sentences are grammatically correct. In addition, many existing summarization systems do not have a function that takes into account keywords that the user wants to include or exclude, nor a function that visualizes the area of the source text from which the summary is copied or the area in the summary that is newly generated. As a result, with existing summarization systems users cannot add constraints to the summary content or efficiently correct grammatical errors in the summary, making them inconvenient to use in actual operation.

To solve the above issues, NTT DOCOMO has developed a multifunctional automatic summarization AI system that creates summaries in accordance with the user's intention. This system has three features. First, the character length control function creates summaries with a length of approximately 70 to 100% of the number of characters specified by the user. Compared with conventional methods, the precision of character length control and the quality of grammar of the summary have been improved. Second, our system has a hint

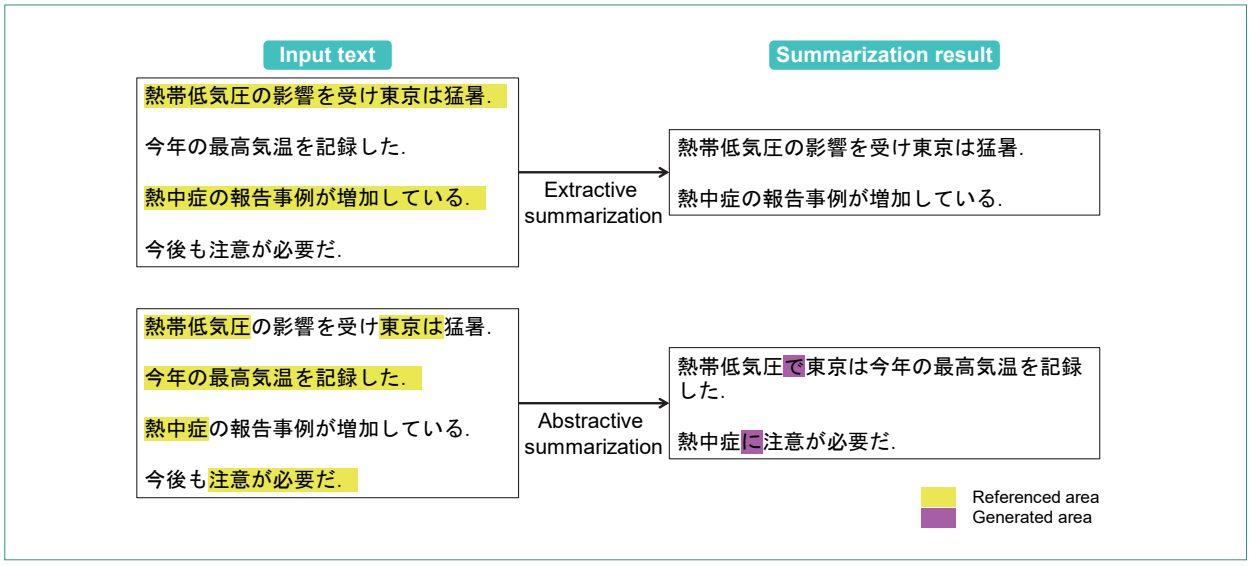


Figure 1 Examples of extractive summarization and abstractive summarization

function that allows users to set keywords to be included or excluded in the summary in order to efficiently generate summaries that meets the user's purpose. The system also has a title function that generates a summary that includes the content of the news article's title. Third, the system has the visualization function that highlights the areas in the source text that are referenced to generate summarization and the locations of newly generated text, which helps users to correct the summary more efficiently.

This system can automatically generate optimal summaries in terms of the number of characters and the content and allow the user to efficiently revise the generated summary when the summary includes errors. Compared with manual summarization and conventional summarization systems, this system can reduce the time required for summarization and addresses the shortage of human resources.

In this article, we describe the prominent functions of our automatic summarization AI system, techniques for improving the system's abstractive summarization performance, and the results of evaluation.

2. NTT DOCOMO's Automatic Summarization AI System

2.1 Overview of System

NTT DOCOMO's automatic summarization AI system is composed of two systems: an extractive summarization system and an abstractive summarization system (**Figure 2**). Both systems use deep

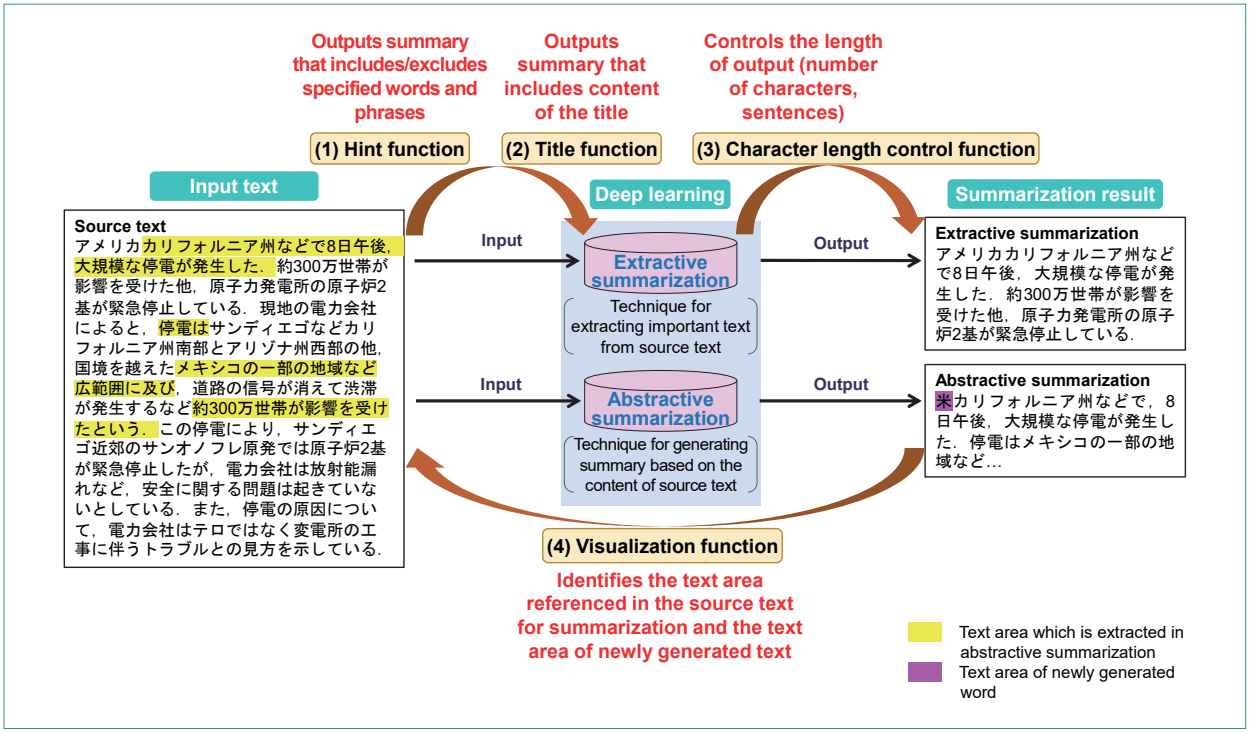


Figure 2 Overview of NTT DOCOMO's automatic summarization AI system

learning*1 and are equipped with the following functions.

(1) Hint function: Adds restrictive conditions to the content of a summary through keywords and phrases specified by the user
(2) Title function: Generates summary that includes the content of the title
(3) Character length control function: Controls the number of characters and sentences in a summary
(4) Visualization function: Allows the user to visually check source text that were referenced to generate the summary and positions of newly generated text.

2.2 Hint Function

The hint function allows users to add restrictions to the content of a summary by specifying keywords or phrases they want to include or exclude. For a news article, the information that different users want to include in the summary, such as names of persons, places, and company appearing in the article, is different. This function was developed to generate summaries optimized for each user.

The system allows users to input multiple keywords and phrases. **Figure 3** shows an example of how the summary changes based on the keyword that the user wants to include in the summary. **Figure 4** shows an example of how the summary changes when the user wants to exclude the keyword suggested by the hint from the summary. In this way, the system makes it possible to output summaries that are closer to what users intend through input of hints.

2.3 Title Function

The title function can improve the quality of the summary by allowing the user to input the title of source text. The quality of deep learning-based summarization usually depends on the quality and quantity of the training data. When the source text contains content that is not included in the training data, the quality of the summary may decline. We thus developed the title function. This

Source text	Specified character length
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．約300万世帯が影響を受けた他，原子力発電所の原子炉2基が緊急停止している．現地の電力会社によると，停電はサンディエゴなどカリフォルニア州南部とアリゾナ州西部の他，国境を越えたメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(294 characters)	120

Result of summarization without the keyword suggested by the hint
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．停電はメキシコの一部の地域など広範囲に及び，約300万世帯が影響を受けたという．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(118 characters)

Result of summarization including keyword “緊急停止” suggested by the hint
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．(107 characters)

Keyword hint

Figure 3 Example of using hint function to produce abstractive summarization that includes hint

*1 Deep learning: A type of machine learning that uses multilayer neural networks.

function improves the quality of summaries by using the title as a hint, which extracts important content from the source text. As shown in **Figure 5**, when the user specifies the title, the system generates a summary that includes keywords in the title.

2.4 Character Length Control Function

The character length control function allows the user to specify the number of characters in a summary. Because the size of a summary display area in a news website or social media is limited,

Source text	Specified character length
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．約300万世帯が影響を受けた他，原子力発電所の原子炉2基が緊急停止している．現地の電力会社によると，停電はサンディエゴなどカリフォルニア州南部とアリゾナ州西部の他，国境を越えたメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(294 characters)	100

Result of summarization without excluding the keyword suggested by the hint
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．停電はメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．(90 characters)

Result of summarization excluding keyword “渋滞” suggested by the hint
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は，安全に関する問題は起きていないとしている．(100 characters)

Excluding keyword hint

Figure 4　Example of using hint function to produce abstractive summarization that excludes hint

Input	Specified character length
【タイトル】 米国などで大規模停電，原子炉2基緊急停止 【原文】 アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．約300万世帯が影響を受けた他，原子力発電所の原子炉2基が緊急停止している．現地の電力会社によると，停電はサンディエゴなどカリフォルニア州南部とアリゾナ州西部の他，国境を越えたメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(294 characters)	100

Result of summarization without specification of title
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．停電はメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．(90 characters)

Result of summarization with specification of title
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は，安全に関する問題は起きていないとしている．(100 characters)

Title-related keywords

Figure 5　Example of abstractive summarization that uses title function

there is a maximum number of the characters in the summary. This system thus is tuned so that the summary is kept within 70 – 100% of the character length specified by the user. As shown in **Figure 6**, when different character lengths are set for the same source text, the system generates summaries that satisfy the length requirement specified by the user. Using this function allows the system to generate summaries of various lengths. Besides the number of characters, the user can also specify the number of sentences.

2.5 Visualization Function

The visualization function allows the user to view the areas of the source text that are referenced to generate a summary and the positions of newly generated text in the summary. Existing automatic summarization AI systems do not show the areas of the source text that are referenced in the summary. As a result, it is time-consuming for the user to compare the source text and the summary to confirm whether or not the summary includes important content from the source text and whether or not the grammar of the generated sentences is correct. We therefore developed the visualization function, which highlights the areas of the source text referenced by the system to generate the summary and the positions of newly generated text. This function thus allows the user to efficiently compare the source text and summary result. As shown in **Figure 7**, the generated summary correctly mapped the phrase "原子炉2基が緊急停止" to the areas in the source text where it occurred even several times. Newly generated text is also correctly visualized.

Source text	Specified character length	Summarization result
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．約300万世帯が影響を受けた他，原子力発電所の原子炉2基が緊急停止している．現地の電力会社によると，停電はサンディエゴなどカリフォルニア州南部とアリゾナ州西部の他，国境を越えたメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(294 characters)	80	アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．停電はメキシコの一部の地域など広範囲に及び，約300万世帯が影響を受けたという．(70 characters)
アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．約300万世帯が影響を受けた他，原子力発電所の原子炉2基が緊急停止している．現地の電力会社によると，停電はサンディエゴなどカリフォルニア州南部とアリゾナ州西部の他，国境を越えたメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．この停電により，サンディエゴ近郊のサンオノフレ原発では原子炉2基が緊急停止したが，電力会社は放射能漏れなど，安全に関する問題は起きていないとしている．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(294 characters)	160	アメリカカリフォルニア州などで8日午後，大規模な停電が発生した．停電はメキシコの一部の地域など広範囲に及び，道路の信号が消えて渋滞が発生するなど約300万世帯が影響を受けたという．また，停電の原因について，電力会社はテロではなく変電所の工事に伴うトラブルとの見方を示している．(136 characters)

Extracted position

Figure 6 Example of abstractive summarization that uses character length control function

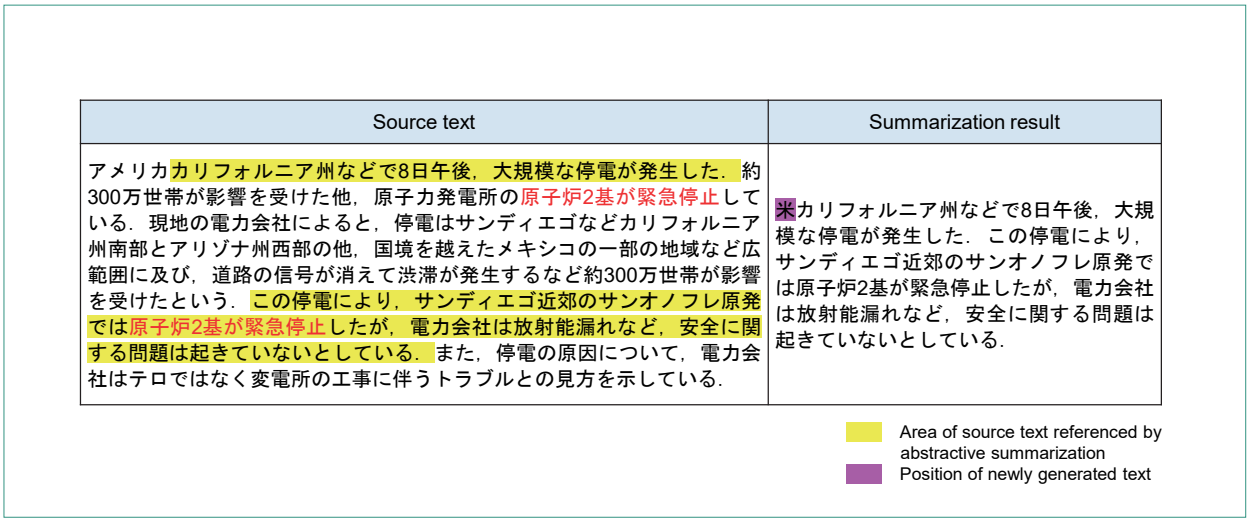


Figure 7　Example of visualization of text areas referenced for summarization and position of newly generated text

3. Methods for Improving Quality of Abstractive Summarization

3.1 Overview

Normally, the training of a deep learning-based summarization model uses a large amount of pair data of source text and human-created summaries. However, manually making summaries is time-consuming and costly.

Instead of requiring a large amount of prepared data, NTT DOCOMO's automatic summarization AI system improves the performance of the summarization model by using summaries containing errors generated mechanically from pair data consisting of source text and summaries as well as large amounts of source text data without correct summaries. It also uses compressed sentences from which extraneous information has been removed. Specifically, we introduced original technologies in the areas of grammar, non-redundancy, fluency, and character length control, as shown in **Figure 8** below. As a result, the system achieved greater performance compared with conventional methods.

(1) We generate sentences with grammatical errors mechanically and use reinforcement learning to reduce grammatical errors.
(2) We generate summaries with redundant content mechanically and deploy contrastive learning*2 to reduce redundancy.
(3) We implement pre-training to improve fluency by using a large amount of source text data.
(4) We deploy a sentence compression model trained using compressed sentences to control character length.

3.2 Grammar

Grammatical mistakes occur especially when summarizing patterns of word combinations that are not included in the training dataset. The cause is the use of "teacher forcing" algorithm [1], a technique widely used when training a deep learning model to generate text. In text generation using deep learning, the last generated word is used to

*2 Contrastive learning: A machine learning technique that increases the accuracy of the model by training the model to learn that the distances of features in similar data are closer than the distances between features in dissimilar data.

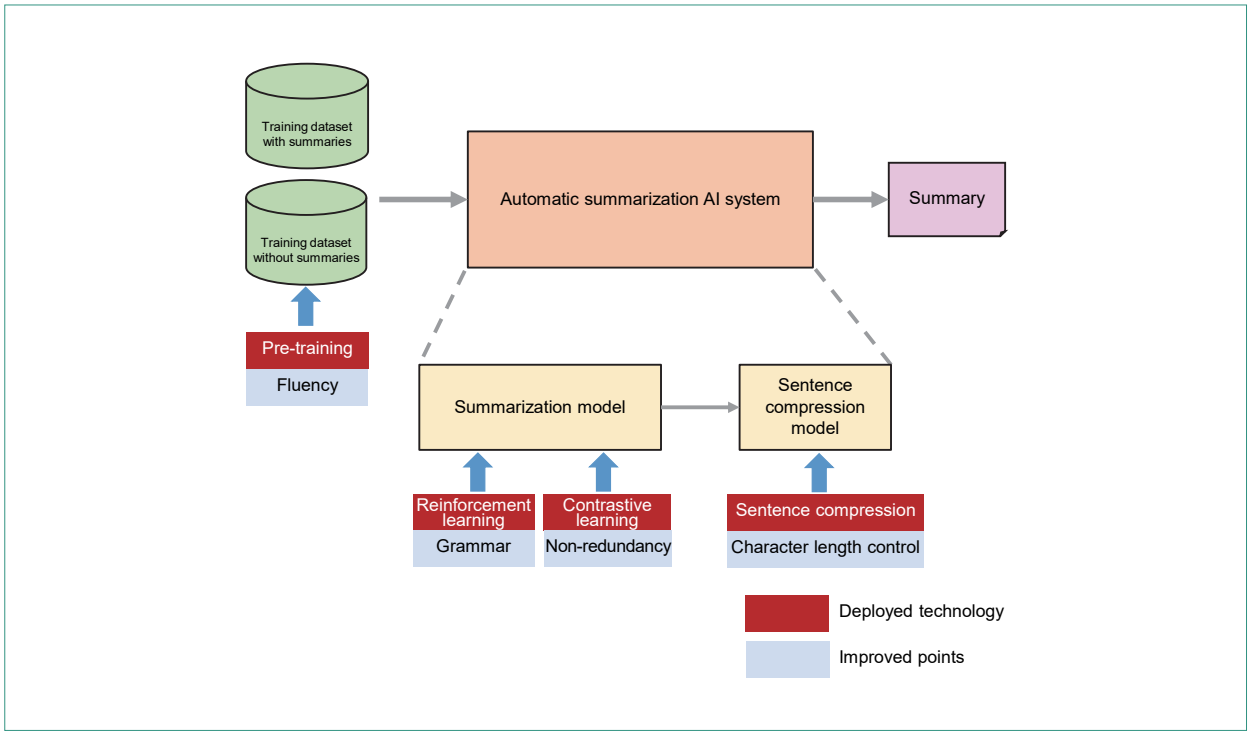


Figure 8 Summarization system's improved areas

generate the next word. Many incorrect words are generated in the initial stage of training, and teacher forcing can make training more efficient by directly inputting correct words into the deep learning decoder*3 instead of using the words previously generated by the model. In such training, because the model learns the next word to be generated for a specific word, word-level rather than entire sentence-level training is performed.

The issue is that even though the correct word is given during training, when generating a text using an unknown sentence, the text is generated using the word generated previously output by the model instead of the correct word. Therefore, once an erroneous word is generated, the model generates text that includes semantic and grammar errors.

To address this issue, NTT DOCOMO's automatic summarization AI system incorporates reinforcement learning. Reinforcement learning trains an intelligent agent through feedback between the agent and its environment, which provides the agent with rewards. While this technique is generally used in fields such as robotic control, in recent years it has also been used in the field of natural language processing [2]. In our system, reinforcement learning is composed of the following elements:

- Agent: The summarization model
- Environment: Discriminator that recognizes whether a given sentence is grammatically correct
- State: Summarization result
- Action: Generation of next word
- Reward: Score of grammatical correctness

*3 Decoder: In a multilayer neural network that takes input text and outputs new text, the part that receives text as input and converts it to feature values is called the Encoder; the part that generates new text by using the feature values output by the Encoder is called the Decoder.

(recognition result of by the discriminator)

By training the summarization model so that it gains the most reward through the action of generating summaries, in other words, by training the model to generate as few grammatical errors as possible, we developed a summarization model that produces fewer grammatical errors compared with conventional methods.

3.3 Non-redundancy

In this article, "redundancy" means that semantically identical content is repeated within the same sentence or across multiple sentences. Deep learning-based text generation models have been observed to have the issue of generating summaries that include redundant content [3].

NTT DOCOMO's automatic summarization AI system thus uses contrastive learning to address this problem and improves text generation performance. Contrastive learning is a technique for improving the accuracy of a model by training it with three datasets: an anchor dataset, positive examples, and negative examples. During training, the model learns features, considering distances between the feature values of the anchor dataset and the positive examples as closer than the distances between the feature values of the anchor dataset and negative examples. In the NTT DOCOMO's automatic summarization AI system, anchors are the summaries generated by the summarization model. Positive examples are correct human-created summaries, and negative examples are incorrect summaries in terms of redundancy, with the same words, phrases, and sentences repeated mechanically. By training the model using negative, erroneous examples in terms of redundancy, we developed a summarization model that generates fewer redundant sentences compared with conventional summarization methods.

3.4 Fluency

In this article, "fluency" expresses the quality of combinations of words and phrases being correct. In actual use, our system may receive as input source text that contains combinations of words that are not included in the training dataset. In such a case, conventional methods may generate a summary with a combination of words that contains wrong usage while being grammatically correct.

To address this issue, NTT DOCOMO's automatic summarization AI system conducts pre-training on word combinations using a large amount of text, resulting in improved fluency in the summarization model.

3.5 Character Length Control

The most common technique for character length control is to choose the output result with the appropriate length based on rules when conducting a beam search[*4] in the process of generating a summary. In such a rule-based method, grammatical correctness and the degree of pertinence are not fully considered when selecting an output with an appropriate length. It is thus easy to output a summary that, while close to the length specified by the user, contains grammatical errors or strays from the main subject of the source text [4].

To address this issue, NTT DOCOMO's character length control function inputs the summary length information to the model as a feature value[*5] in order to optimize the model to simultaneously

*4 Beam search: In this article, refers to the selection of multiple candidate words output by the neural network based on their scores, resulting in several summarization result candidates.

*5 Feature values: Values extracted from data, and given to that data to give their features.

learn grammar, pertinence, and length. In addition, as a post-processing technique, a text compression model is applied to reduce the number of characters in the summary. If the output exceeds the character length specified by the user, the system can compress the summary to the appropriate length.

4. Evaluation of Performance of Abstractive Summarization

4.1 Dataset

We used a dataset of about 180,000 news articles and their human-created summaries, provided by the Nippon Television Network Corporation, as the dataset for training and evaluating our system's abstractive summarization model.

4.2 ROUGE Evaluation

Figure 9 shows how each metric of Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is calculated. ROUGE evaluates the comprehensiveness (recall) of a summarization by comparing with a correct summary (reference). ROUGE is the most widely used metric for evaluating summarization

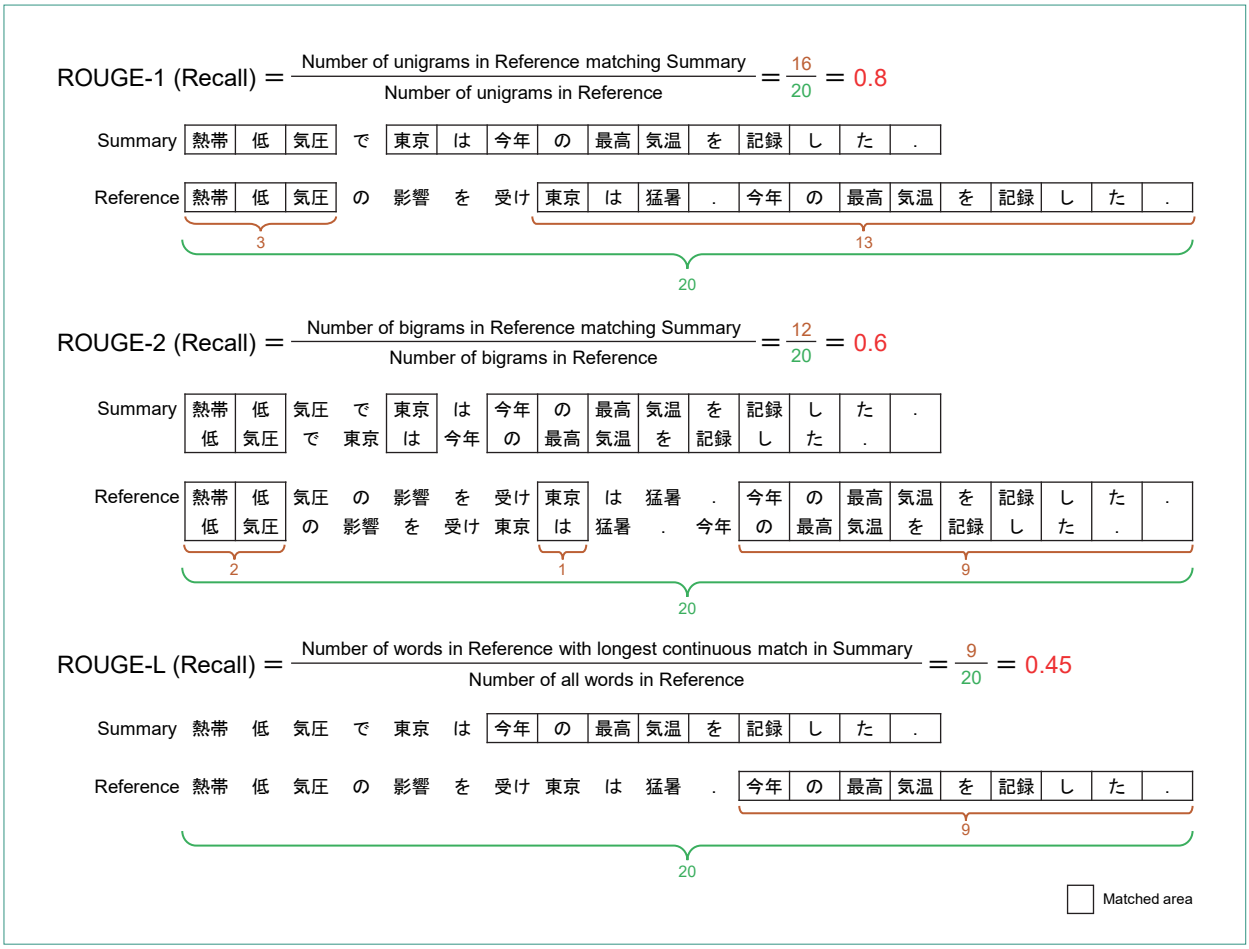


Figure 9 ROUGE metrics calculation methods

by comparing the similarity of text generated by a model with the reference text [5]. ROUGE-1 and ROUGE-2 represent the degree of overlap of unigrams*6 and bigrams*7, respectively, between the generated text and the reference text. ROUGE-L measures the degree of overlap using the length of longest common text fragment. For all ROUGE metrics such as ROUGE-1, ROUGE-2, and ROUGE-L, the higher the score, the greater the degree of overlap between the two texts, and the greater the performance of the text generation model.

Table 1 shows the evaluation results of NTT DOCOMO's summarization system using 3,000 texts from Nippon Television Network Corporation's dataset. We compared our summarization model with Lead-3, which takes the first three sentences of the source text as the summarization result, as well as TextRank [6], SumBasic [7], LSA (Latent Semantic Analysis) [8], Submodular [9], and PGN (Pointer Generator Network) [3]. TextRank, SumBasic, LSA, and Submodular are extractive summarization techniques and extract the three most important sentences in the source text as the summarization result. PGN is an abstractive summarization technique. The summary closest to the character length of the most correct summary in beam search results is evaluated as the final summarization result. Compared to these summarization methods, NTT DOCOMO's automatic summarization AI system achieved higher ROUGE scores, indicating that it generates summaries close to correct, human-created summaries.

4.3 Human Evaluation

One of the disadvantages of ROUGE evaluation described above is that it cannot evaluate grammatical mistakes and semantic redundancies. Thus, a person whose native language is Japanese evaluated summaries using the following four criteria. Each criterion was evaluated on a 4-point scale, with 4 being the highest score.

(1) Grammar: Few grammatical mistakes in the summary
(2) Pertinence: The generated summary covers the main content of the source text
(3) Non-redundancy: Semantically identical words, phrases, and sentences are not repeated in the summary

Table 1 Results of ROUGE evaluation

		ROUGE-1	ROUGE-2	ROUGE-L
Extractive summarization	Lead-3	74.46	63.89	72.48
	TextRank	64.06	50.07	60.16
	SumBasic	64.49	49.18	58.38
	LSA	62.28	46.48	56.85
	Submodular	55.41	36.91	47.41
Abstractive summarization	PGN	79.25	70.36	77.45
	NTT DOCOMO	84.49	76.47	81.80

*6 Unigram: A string of n consecutive words is called an n-gram. When n is 1, the string consists of only one word.
*7 Bigram: A string of two consecutive words.

(4) Fluency: The flow from word to word and sentence to sentence is smooth in the generated summary

Table 2 shows the average scores of human evaluation of 100 summaries produced from the evaluation dataset by NTT DOCOMO's automatic summarization AI system and PGN. An additional score on a 4-point scale is calculated for each summary's character length; the criteria for each score, based on attainment of the specified character length, are given in **Table 3**. This score is calculated for each generated summary, and the average of all the scores is added to the last column "Character length" in Table 2. The results of Table 2 show that NTT DOCOMO's automatic summarization AI system scores higher than the PGN method in all evaluation metrics, indicating that it generates higher-quality summaries. Note that our system's automatic summarization takes about 1 second for extractive summarization and 10 seconds for abstractive summarization. These performances are much faster than human summarization, which requires several minutes.

5. Conclusion

In this article, we described the functions, techniques, and performance of NTT DOCOMO's automatic summarization AI system. NTT DOCOMO has developed the following functions to realize a system that easily outputs summaries according to the user's intentions: a hint function, which allows the user to specify hints when summarizing;

Table 2 Results of human evaluation

	Grammar	Subject matter pertinence	Non-redundancy	Fluency	Number of characters of summary
PGN	2.82	2.40	3.89	3.06	3.25
NTT DOCOMO	3.85	3.53	3.92	3.84	3.89

Table 3 Score related to number of characters of generated summary

Range of the number of characters in summary	Score
$0.7\times L \leqq S \leqq 1.0\times L$	4
$0.6\times L \leqq S < 0.7\times L$ or $1.0\times L < S \leqq 1.1\times L$	3
$0.5\times L \leqq S < 0.6\times L$ or $1.1\times L < S \leqq 1.2\times L$	2
$S < 0.5\times L$ or $1.2\times L < S$	1

L: The number of characters specified by user
S: Character count of summary generated by the summarization model

a title function, which allows the title to be used in summarizing; a character length control function, which controls the number of characters and sentences in a summary; and a visualization function, which helps the user to view the areas of text referenced in the source text and the positions of newly generated text. To solve issues related to grammar, non-redundancy, fluency, and summary character length control, we deployed techniques such as reinforcement learning, contrastive learning, and pre-training; these techniques led to improvement of performance. As the results of performance evaluation show, our automatic summarization AI system greatly outperformed conventional methods in ROUGE metrics and human evaluation. NTT DOCOMO's automatic summarization AI system can reduce the time required for summarization and address the shortage of human resources. We will continue to improve the performance of existing functions and develop new functions to respond to issues experienced in actual service and realize an even higher-performing automatic summarization AI system.

REFERENCES

[1] R. J. Williams and D. Zipser: "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks," Neural Comput., Vol.1, No.2, pp.270–280, 1989 (doi: 10.1162/neco.1989.1.2.270).

[2] L. Yu, W. Zhang, J. Wang and Y. Yu: "SeqGAN: Sequence generative adversarial nets with policy gradient," 31st AAAI Conf. Artif. on Intell., pp.2852–2858, 2017.

[3] A. See, P. J. Liu and C. D. Manning: "Get To The Point: Summarization with Pointer-Generator Networks," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.), Vol.1, pp.1073–1083, 2017 (doi: 10.18653/v1/P17-1099).

[4] B. Eikema and W. Aziz: "Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation," Proc. of the 28th International Conference on Computational Linguistics, pp.4506–4520, 2021 (doi: 10.18653/v1/2020.coling-main.398).

[5] C.-Y. Lin: "Looking for a Few Good Metrics: ROUGE and its Evaluation," NTCIR Work., pp.1–8, Jun. 2004.

[6] R. Mihalcea and P. Tarau: "TextRank: Bringing Order into Text," Proc. of 2004 Conf. Empir. Methods Nat. Lang. Process., pp.404–411, 2004.

[7] L. Vanderwende, H. Suzuki, C. Brockett and A. Nenkova: "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," Inf. Process. Manag., Vol.43, No.6, pp.1606–1618, Nov. 2007 (doi: 10.1016/j.ipm.2007.01.023).

[8] M. G. Ozsoy, F. N. Alpaslan and I. Cicekli: "Text summarization using latent semantic analysis," J. Inf. Sci., Vol.37, No.4, pp.405–417, Aug. 2011 (doi: 10.1177/0165551511408848).

[9] H. Lin and J. Bilmes: "A class of submodular functions for document summarization," Proc. of 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., Vol.1, pp.510–520, Jun. 2011.

Technology Reports (Special Articles) Behavior Prediction Recommendations RNN

Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

A Recommendation Engine Using Time Series Prediction Models of User Behavior

Service Innovation Department Taku Ito Yuri Sasaki Shigeki Tanaka

The provision of a recommendation function for products and content in B2C services on the web has become a common feature in recent years. Recommendation engines come in various types, but most are based on popularity ranking, which means that they have not been able to truly understand the context of service use by a particular user. To present recommendations that attract the user's interest, a recommendation engine must be able to understand user context and predict content that the user should find interesting. With the aim of raising the probability of purchases, NTT DOCOMO has developed a recommendation engine that interprets behavior based on time series data and applies a deep learning algorithm to make predictions. This engine has been applied to NTT DOCOMO services enabling the provision of highly accurate recommendations.

1. Introduction

It has become common in Electronic-Commerce (EC) services*1 and Business-to-Consumer (B2C) services to provide product recommendations*2. A recommendation function is also being provided in many NTT DOCOMO services as "recommendations for the user." A wide range of recommendation techniques have been proposed from simple techniques based on popularity ranking to those

*1 EC services: Services that enable the buying and selling of goods on the Internet.

using machine learning*3, and in just the last few years, an easy-to-use machine learning library*4 has appeared on the market and recommendation services using cloud services*5 have begun to be provided.

Recommendation functions have been provided for a number of years, but at present, many recommendation systems are based on popularity ranking*6. This approach determines the preferences of a majority of users and can therefore achieve a certain Click-Through Rate (CTR)*7, but it is not without its problems. For example, since the same products are recommended to different users in such a system, those recommendations have no effect on users who have no interest in popular content, and given that the content presented is fixed, many other types of products go without recommendations making it difficult for users to discover new products. Additionally, since the procedure typically followed by a user is to click on desirable products during a sequence of web movements, time series data that can reveal the sort of content the user has just clicked on is more important than interests or preferences in a recommendation service. For example, given a user searching for food on a portal site, it would be desirable to present food against the background of the user's web movements without regard to everyday interests or preferences. If the user has just clicked on "fresh foods," the user can then be presented with fresh foods thereby enabling content desired by the user to be recommended.

Against the above background, NTT DOCOMO has applied a Recurrent Neural Network (RNN) as a time series prediction algorithm to the generation of recommendations so that long-term and short-term shifts in user interests can be determined. This technology mainly makes recommendations in genres that the user has recently showed an interest in but can also make recommendations in genres that the user has regularly been interested in. In this way, recommendations can be presented tailored to personal conditions, which can help improve CTR. In this article, we describe recommendation algorithms that have actually been applied and examine their effectiveness in certain NTT DOCOMO services.

2. Using RNN to Make Recommendations

2.1 RNN Overview

An RNN, which is also called a recursive type of neural network*8, is a neural network model designed to recognize patterns in sequential data such as time series data*9 [1]. The conventional neural network model features a fixed-length input layer and output layer plus middle layers between those two layers (**Figure 1**). In an image recognition task, for example, the actual image can be divided into units of pixels so that pixel values are treated as the input layer while categories as targets of classification are output on the output layer. In this example, the input layer and output layer consists of a number of pixels and number of classification patterns, respectively, with each taking on a single value. This means that each node*10 must be of fixed length.

On the other hand, actual data cannot necessarily be given in a fixed length. For example, in a task like document prediction, a document is given as input, but since a document is not of fixed

*2 Recommendations: The process of recommending products and content tailored to the user.
*3 Machine learning: Computer algorithms for learning patterns based on input data and executing various types of tasks.
*4 Library: A collection of high-versatility programs in a reusable form.
*5 Cloud services: The provision via the network of services running on remote servers instead of on the user's computer.
*6 Ranking: The listing of products and content in the order of most clicked or purchased over the entire service. It is one of the most basic types of recommendation algorithms.

length as in the case of an image, it cannot be expressed in terms of the above model. An improved model that can handle variable length data is RNN, which handles the points making up variable length data as a time series and stores variables called "states" within the model (in middle layers) (**Figure 2**). The state changes according to input data and that state propagates to the next time series step. In addition, a value corresponding to that state is presented in the output layer. In this way, a model in

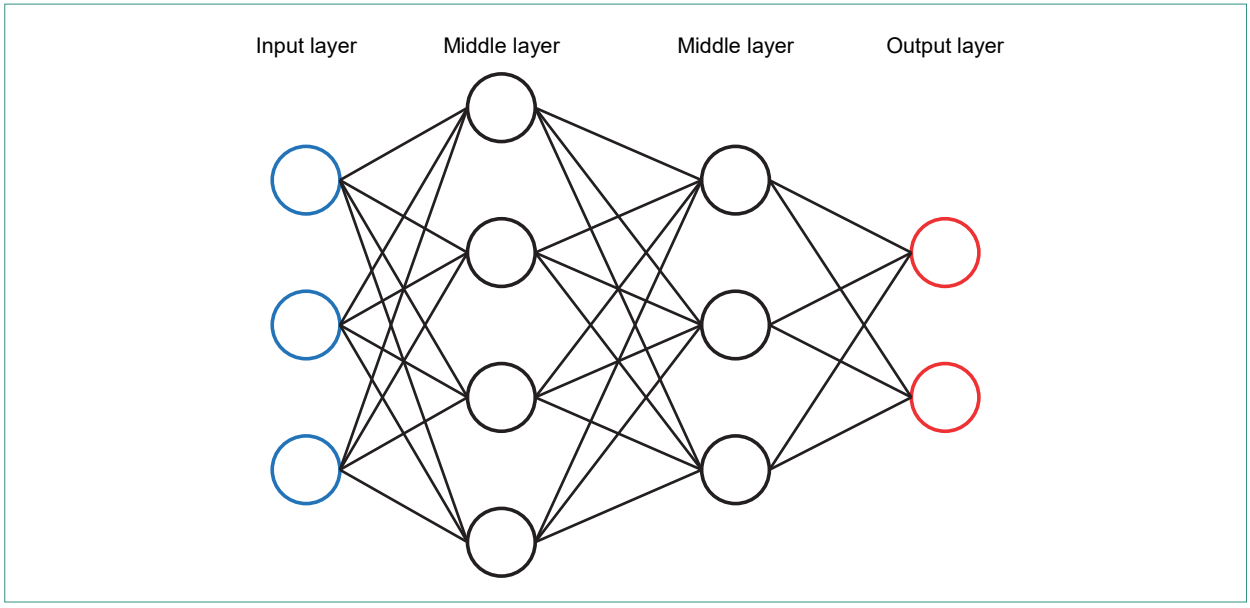


Figure 1 Ordinary neural network

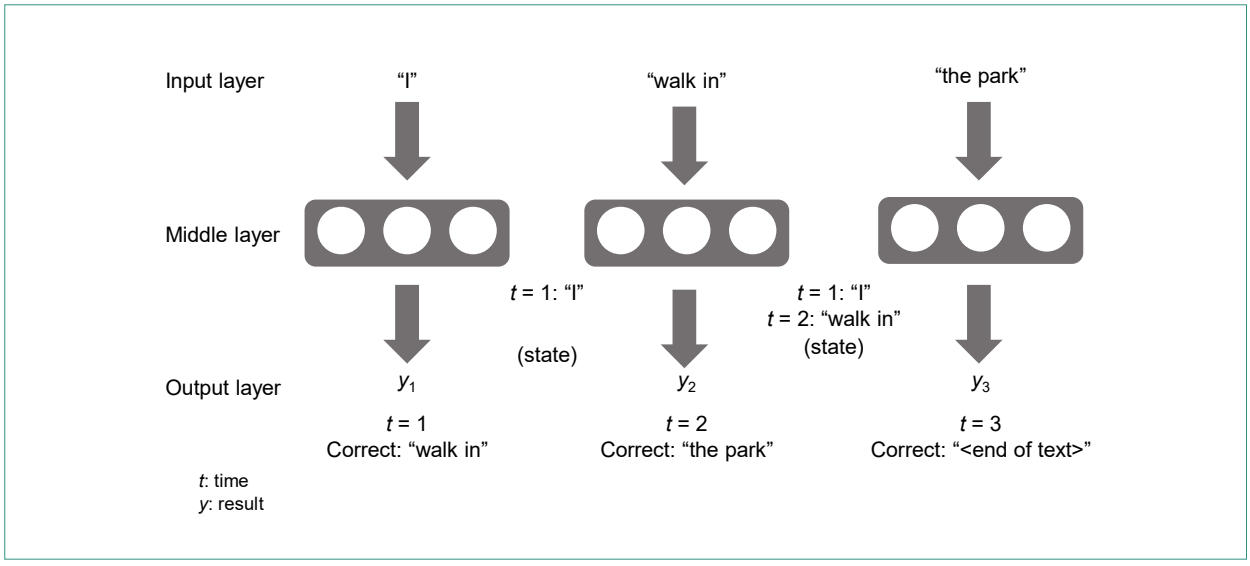


Figure 2 Schematic diagram of RNN

*7 CTR: The ratio of the number of times the user has actually clicked on content to the number of times content has been displayed by a recommendation function.

*8 Neural network: A representation of the neural network inside the human brain by a numerical model consisting of an input layer, middle layers, and output layer.

*9 Time series data: Data containing information on the temporal change in values.

*10 Node: A network point that propagates a value received from input.

which states propagate along a time series makes it possible to apply a neural network to target applications that take variable length data as input.

2.2 Application of RNN to Recommendations

The RNN described above is often used in tasks targeting text such as document recognition*11 and machine translation*12 and tasks targeting speech such as speech recognition. Here, by having a system learn about the data within a network in the manner of "what kind of output value should be given for a given input value," a desired output value with respect to new text or speech can be obtained. Because of this characteristic, RNN can also be used in time series prediction. Training a system in which input and output values are of the same type enables the construction of a model that can predict the next value to come given a specific input [2].

Recent years have seen an increasing number of RNN time series prediction tasks being applied to recommendation systems. In such a system, each input value can be defined as an instance of user behavior such as a product click or video view and an output value can likewise be given as user behavior. In this way, given a user with an interest in a certain product or video, the transition in a user's interest (state), that is, what kind of product or video would that user have an interest in next, can be predicted (**Figure 3**). In addition, since it then becomes possible to make recommendations in line with the user's state at that time, inputting the latest instance of user behavior into the model enables the real-time provision of updated recommendations at all times.

3. Proposed Techniques

3.1 Overview

On applying RNN to the process of making recommendations, we have been making several improvements tailored to the service domain and devising measures to output recommendations applicable to the user's recent interests. In particular, we constructed (1) a model for making recommendations that combines a long-term prediction model and short-term prediction model using hierarchical

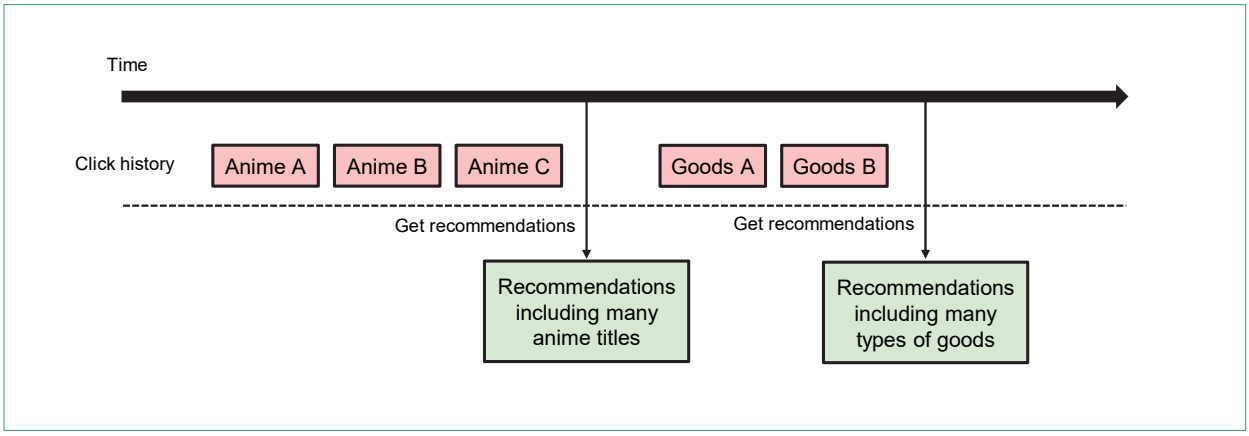


Figure 3 Overview of using RNN for recommendations

*11 Document recognition: A task that determines the category that a certain document belongs to.
*12 Machine translation: A task that automatically translates a document in a certain language to a document in another language using a computer.

RNN and (2) a model for making recommendations that combines multiple RNNs using different information. These models are explained below.

3.2 Long-term/short-term Prediction Model Using Hierarchical RNN

We can consider that user interests undergo both long-term transitions and short-term transitions as opposed to changes over a fixed period [3]. For example, in a system that recommends smartphone applications (apps) based on the user's app installation history, the case of searching for a Social Networking Service (SNS) app on day X and the case of searching for a video delivery app on day X+1 can be considered. It is therefore desirable that SNS apps be recommended on day X and that video delivery apps be recommended on day X+1. However, a simple RNN cannot consider any differences between day X and day X+1, so the possibility exists that SNS apps will be recommended on day X+1 too on the basis of the user's history on day X. To prevent this from happening, it is necessary to recommend SNS apps during day X based on the user's installation history of that day, and when crossing into a new day, to predict a long-term transition in user interests and make recommendations accordingly. Here, a long-term transition in interests refers, for example, to the searching for a video delivery app on the day following a search for an SNS app.

We propose a hierarchical RNN to achieve recommendations that grasp the user's interests over the long term and short term as described above (**Figure 4**). Specifically, the above problem can be

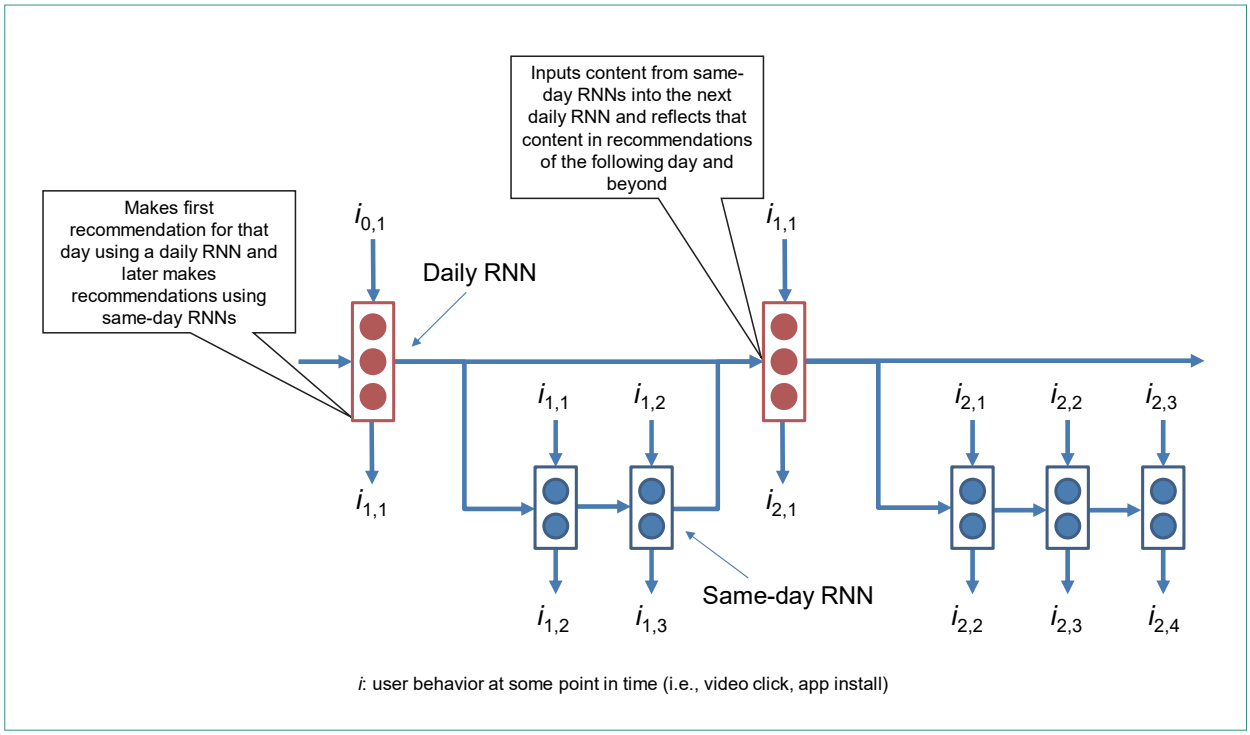


Figure 4 Schematic diagram of hierarchical RNN

solved by dividing the process into a daily RNN, which determines the content for making recommendations from user history up to the previous day, and same-day RNNs, which determine the content for making recommendations from user history of the previous day and current day.

3.3 Recommendation Model that Combines Multiple RNNs Using Different Information

Recommending certain products or videos must be based on the content of such items. On the other hand, the data handled by RNNs consist of numerical values, so it is necessary to handle products, videos, and other goods as numerical data. In general, content IDs are allocated to products, videos, etc., which enables an RNN to predict the time series transition of that ID-based numerical data. With this technique, predictions are made in units of content IDs, but if the same type of content does not exist in the user's past history, no predictions can be made. However, a similar type of time series transition can be considered even in the case of a different content ID as long as similar content exists, so the content (category) of products and video must be taken into account. For example, we can consider that a user who has clicked on a product in the eating-and-drinking category has a high probability of clicking on another product in the same category. It is therefore desirable that predictions be made not only for a time series in units of content IDs but also for a time series in units of categories.

To achieve recommendations in the manner described above, we propose a recommendation model that combines multiple RNNs (**Figure 5**). Combining RNNs that make time series predictions while making data on different layers independent of each other makes it possible to achieve both detailed recommendations in units of content and

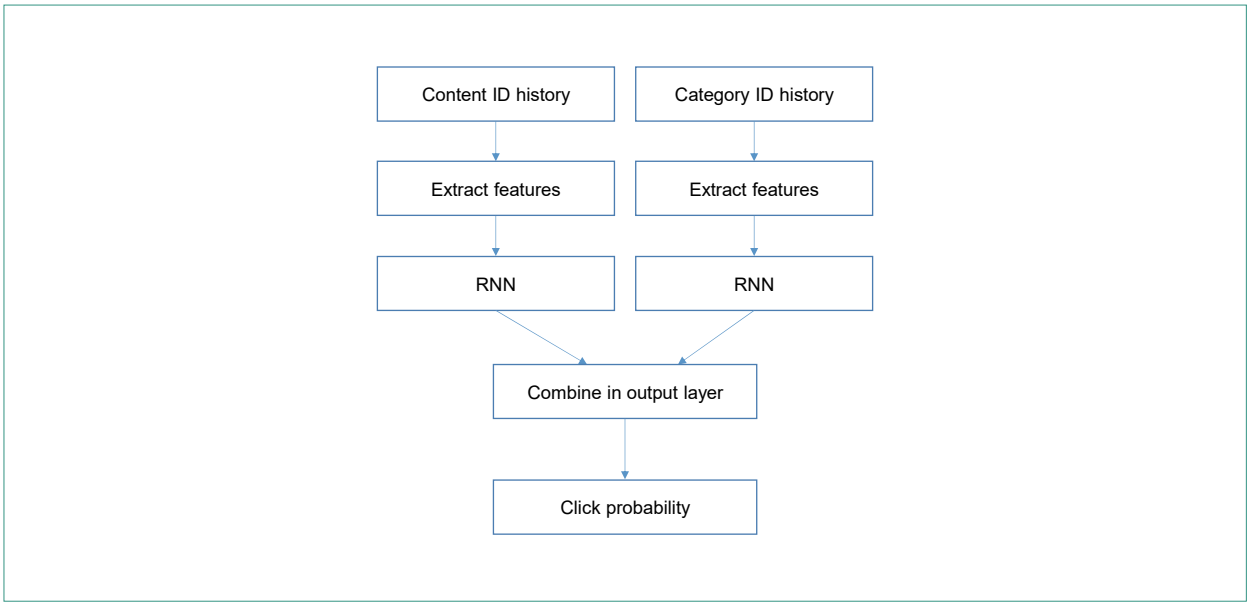


Figure 5 Schematic diagram of multiple RNNs

general recommendations in units of categories thereby improving both accuracy and coverage[*13].

4. Application to NTT DOCOMO Services

We tested the effectiveness of applying the above techniques to actual NTT DOCOMO services, namely, "App Recommender" that recommends apps to the user and the NTT DOCOMO-managed portal site called "dmarket" that recommends products of B2C services. We applied the "recommendation model using hierarchical RNN" to App Recommender and the "recommendation model using multiple RNNs" to dmarket. The following gives an overview of each service and presents the results of testing the effectiveness of each technique.

4.1 App Recommender

The service screen is shown in **Figure 6**. This service presents apps thought to be useful for the user based on the user's history of installing apps. With the aim of getting the user to install more of the displayed apps, we tested the effectiveness of the recommendation model using the hierarchical RNN described above with actual users. Specifically, given the proposed technique and an existing recommendation algorithm using user features[*14] based on history, we conducted an A/B test[*15] of these two methods from September 2020. The proposed technique achieved an install rate[*16] 2.7 times that of the existing algorithm. The reason for this high install rate can be given as follows. Taking into account time series data, a user who has just installed, for example, Twitter and a camera app in succession could be recommended Instagram.

Figure 6 Service screen of App Recommender service

In other words, the characteristics of apps desired by a user could be inferred from immediately previous user behavior.

4.2 dmarket

The service screen is shown in **Figure 7**. As shown here, this service displays content recommended for the user in a cross-service manner in domains classified by 16 tabs such as "shopping" and "gourmet." The objective here is to display recommendations that will induce the user to click on content of interest spanning multiple services. Additionally, the recommended content can be varied each time it is displayed during web movements to continuously attract the user's interest. We conducted a test using actual users on the effectiveness of recommendations that combine multiple RNNs as described above. Specifically, on conducting an A/B test against this technique and the

*13 Coverage: The ratio of the total number of content items recommended to all users to the total number of all content items. High coverage indicates that a wide range of content is being recommended.
*14 Features: Quantities (numerical values) extracted from data and characterizing that data.
*15 A/B test: A test that compares two algorithms to determine which is more effective.
*16 Install rate: The ratio of apps that the user has actually installed to all apps presented by a recommendation function.

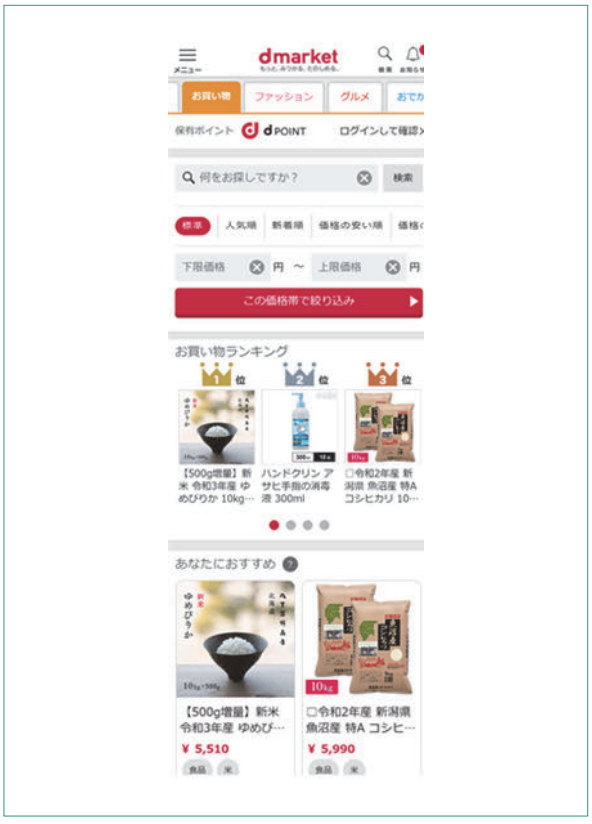


Figure 7　Service screen of dmarket service

existing recommendation algorithm based on popularity ranking, we achieved a 2.0% improvement in CTR. We consider the reason for this improvement in CTR is that the proposed technique could extract common elements from items of content belonging to different services. For example, many food items can be recommended by popularity ranking on the gourmet tab, but for a user that frequently uses dmagazine*17, many gourmet magazines can be recommended as well. In short, it became possible with the proposed technique to recommend services applicable to that user even for the same gourmet tab.

5. Conclusion

This article described recommendation algorithms as an extension of RNN. It described, in particular, recommendations using hierarchical RNN for determining shifts in user interests in both the long term and short term tailored to service domain characteristics and recommendations that combine multiple RNNs using multiple features of content. Both algorithms were tested for their effectiveness in actual NTT DOCOMO services and were found to lead to improved accuracy compared with existing algorithms. At NTT DOCOMO, we aim to pursue the latest technologies in recommendation algorithms and to promote initiatives that raise the value of service provision.

REFERENCES

[1] J. Li, L. Deng, R. H.-Umbach and Y. Gong: "Robust Automatic Speech Recognition: A Bridge to Practical Applications," Academic Press, Oct. 2015.

[2] D. Wierstra, J. Schmidhuber and F. Gomez: "Evolino: Hybrid Neuroevolution/Optimal Linear Search for Sequence Learning," Proc. of the 19th International Joint Conference on Artificial Intelligence (IJCAI), Edinburgh, pp.853–858, Jul. 2005.

[3] B. Hidasi, A. Karatzoglou, L. Baltrunas and D. Tikk: "Session-based recommendations with recurrent neural networks," arXiv preprint arXiv: 1511.06939, Nov. 2015.

*17 dmagazine: A subscription-based magazine delivery service provided by NTT DOCOMO.

Technology Reports (Special Articles) AI Image Recognition Deep Learning

Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

Providing Image Recognition AI via the DOCOMO Image Recognition Platform

Service Innovation Department Toshiki Sakai Motoki Iwata

In recent years, AI technology has become widespread, and in the field of image recognition, it is being used to replace, automate, and save human labor. However, its introduction has necessitated the preparation of high-performance server environments, the installation of various software packages, and the creation of AI development environments. For this reason, we have developed the DOCOMO Image Recognition Platform to facilitate the development and deployment of diverse image recognition AI systems. This has made it possible for AI users to introduce image recognition AI technology simply by preparing data for image recognition AI instead of having to prepare their own AI development and operation environments.

1. Introduction

In recent years, AI technology based on deep learning*1 has made remarkable advances and has become very popular. Particularly strong progress has been made in industrial applications of image recognition AI*2. For example, it is being used to support or even replace tasks that have hitherto been done by humans, such as assisting with product inspections in factories, detecting people and vehicles in security camera images, or supporting diagnostic imaging in medical care. However, when image recognition AI is used for the streamlining, automation and Digital Transformation (DX)*3 of

*1 Deep learning: A machine learning method that can learn more complex concepts and perform judgments and estimations by using a more complex form of neural network, which is designed to imitate human neural processing mechanisms.
*2 Image recognition AI: A form of AI that takes images as its input and uses them to generate results by making judgments, performing estimates, and so on.

existing jobs, it is seldom available in a form that can directly solve the issues faced by each company. For this reason, each company must first prepare its own collection of images and/or video data together with annotations that record the answers corresponding to the type of image recognition processing the AI is needed to perform on the data, and must then perform training to develop its own image recognition AI, followed by deployment*4 in order to make the trained image recognition AI usable. Deployment involves setting up this image recognition AI on a server and building an Application Programming Interface (API)*5 that inputs images and outputs results as text or the like. However, for AI users to perform this training and deployment by themselves, there are several hurdles they have to tackle, including understanding the framework*6 used for working with deep learning, performing high-speed training, and preparing hardware for recognition.

To make things easier, services are available that facilitate the training and deployment of image recognition on the cloud. Examples include Amazon Rekognition Custom Labels [1], and AutoML Vision [2]. In 2020, we released the DOCOMO Image Recognition Platform [3] to make it easier for more users to access the image recognition*7 technology we developed in-house.

When providing functions for training and deploying image recognition AI, since each user requires different image recognition functions, a mechanism is required for providing these functions more efficiently. This article describes the mechanism used to provide multiple image recognition functions on the DOCOMO Image Recognition Platform. The DOCOMO Image Recognition Platform also provides a more secure image recognition environment by performing image recognition within the DOCOMO's closed network, as will be explained below.

2. Overview of the DOCOMO Image Recognition Platform

The DOCOMO Image Recognition Platform provides functions for training image recognition AI in the cloud and deploying it in a state where inference using image recognition AI can be performed.

1) Image Recognition Functions

Figure 1 shows the image recognition functions provided by the DOCOMO Image Recognition Platform: (1) object detection (find a specific object in an image and estimate its coordinates within the image), (2) generic object recognition (classify objects into categories based on features of the objects themselves, their surroundings, the scene, and the image as a whole), (3) character recognition (recognize text characters in an image), (4) similar image search (search for images similar to a given image), (5) specific object recognition (identify a specific object by matching an image of this object with multiple pre-prepared images of objects), and (6) pose estimation (estimate the position of a person's skeleton and joints in an image).

2) Assumed Use Cases

Figure 2 shows the assumed use cases of each image recognition function. Object detection can be used to detect weed growth in drone images and people and cars in surveillance camera images.

*3 DX: The changes that digital technology causes or influences in all aspects of human life.

*4 Deployment: Installing applications by placing them in their execution environments.

*5 API: An interface that enables the functions of software to be used by other programs.

*6 Framework: Software that encompasses functionality and control structures generally required for software in a given domain. In contrast to a library in which the developer calls individual functions, code in the framework handles overall control and calls individual functions added by the developer.

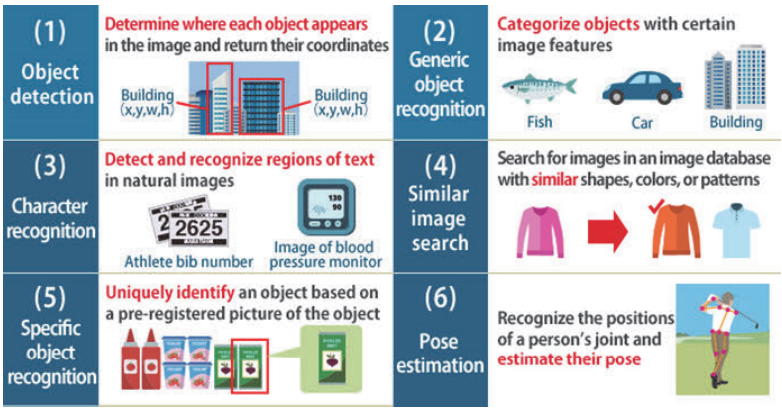


Figure 1 Image recognition functions provided by the DOCOMO Image Recognition Platform

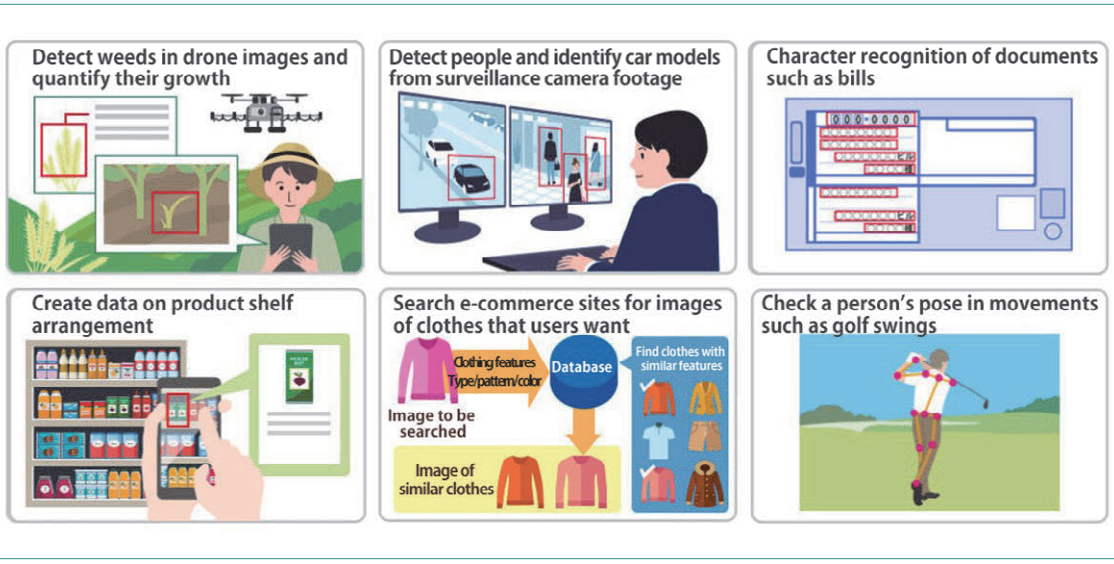


Figure 2 Assumed use cases of each image recognition function

By combining generic object recognition with the results of detecting areas that contain plants or cars, it is also possible to determine the amount of plant growth and the types of cars. Another possible use is to combine object detection and character recognition to read characters from bills and documents, or to combine object detection and specific object recognition so that images of product shelves can be converted into data on what products are stored where.

In addition, similar image search can be applied to images of fashion items in order to find similar items, and pose estimation can be used to check the form of actions performed by sports players.

*7 Image recognition: Technology that uses image processing and machine learning (see *8) to enable machines to understand images and extract meaning from them.

3) Functional Configuration

The configuration of the DOCOMO Image Recognition Platform is shown in **Figure 3**. For object detection and generic object recognition (category classification), the platform provides both training and inference functions as custom training models that can implement tailor-made AI based on data provided by the user. In addition, for similar image search and specific object recognition, it provides a user dictionary creation function that enables searching and recognition based on the user's own image data. For object detection, generic object recognition (category classification), character recognition, and pose estimation, NTT DOCOMO also provides pre-trained image recognition AI (common pre-trained model) that was trained by NTT DOCOMO so that users don't have to prepare their own training data to make inferences.

4) Provision of Website/web Console

The DOCOMO Image Recognition Platform provides users with a website/web console. By accessing this console, users can train their AI, check the results of this training (evaluation), and deploy the image recognition AI that they have created and trained. Users can input recognition requests (inference requests) to the deployed AI through a WebAPI interface. When image data is input to the WebAPI, the recognition results are returned in text format.

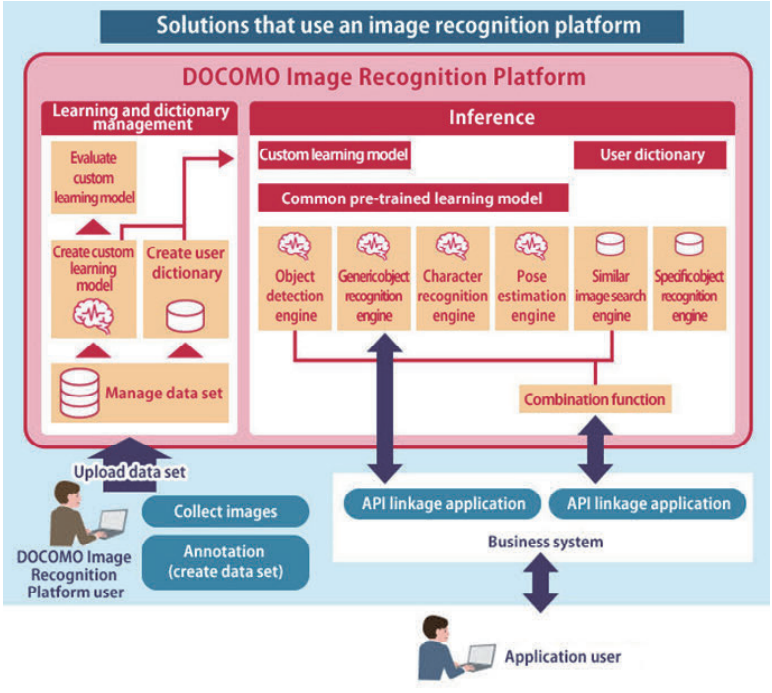


Figure 3 Functioning of the DOCOMO Image Recognition Platform

3. Two Initiatives on the DOCOMO Image Recognition Platform

We are working to make the DOCOMO Image Recognition Platform more convenient for users through two initiatives. The first is to containerize the image recognition functions so that they can be provided, developed and updated more quickly. This makes it possible to provide a full lineup of image recognition functions, enabling the provision of functions that can solve users' problems. The second is to build the DOCOMO Image Recognition Platform on a data center within the DOCOMO's closed network, so that it can be accessed without sending anything over the Internet. This makes it possible to provide secure image recognition AI.

3.1 WebAPI for Image Recognition AI Using Containers

As mentioned above, the DOCOMO Image Recognition Platform provides multiple image recognition functions. The image recognition functions required by users vary depending on the problem they want to solve and are expected to become more varied in the future. Advances in image recognition AI technology are being made every day, and to provide users with better accuracy and higher speeds, the image recognition functions that have already been deployed must be regularly updated. Therefore, to speed up the provision, development and updating of each image recognition function in the DOCOMO Image Recognition Platform, the functions are virtualized using containers and the containerized functions are modularized and shared, as illustrated in **Figure 4**.

Container-based virtualization is a technology whereby applications such as image recognition functions are combined with the libraries needed to run them in the form of packages called containers, allowing these applications to run with minimal

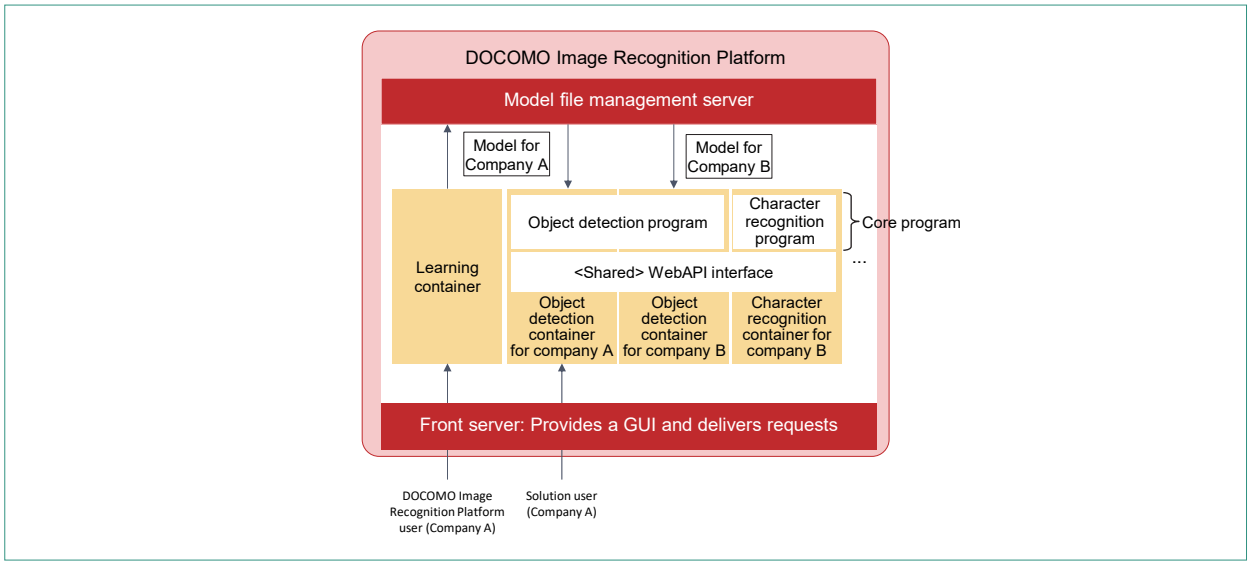


Figure 4 Using containerization to provide image recognition functions

dependence on the server's OS/environment. In the DOCOMO Image Recognition Platform, each image recognition function's core program and its WebAPI interface are combined in a single container. In this way, it is possible to add new image recognition functions, update existing image recognition functions, and augment the recognition resources (changing them so that taining and inference can be performed from more images) by adding/removing containers within the DOCOMO Image Recognition Platform.

Furthermore, within an image recognition container, the program that handles the abovementioned interface is separate from the core program for image recognition, and by using a shared program with a standardized interface, less work needs to be done to develop new image recognition functions. The input/output format of the interface between the interface program and the core program can be flexibly changed from within each core program as shown in **Table 1** to ensure it has the flexibility to enable the expansion of image recognition functions.

In addition, the image recognition models trained for each image recognition AI user are stored independently outside the container. This makes it possible to augment only the resources of a specific user-generated image recognition AI, or to replace just a single container and inherit its model when a container is updated.

3.2 Secure Image Recognition Using the DOCOMO Open Innovation Cloud

The use of image recognition may require the input of sensitive image data. To prevent images from being released to the outside world when performing image recognition, it is important to consider the security of the transmission paths used to transfer these images, and of the servers where the image recognition processing is performed.

When performing image recognition, the image recognition system can be implemented in the cloud or in its own data center to consolidate the processing functions. This makes it possible to increase the utilization of computational resources in the cloud or data center so that image recognition processing can be performed by making efficient use of these resources. When doing so, the risk can be reduced by equipping the system with appropriate security measures. However, even in this case,

Table 1 Division of roles between the image recognition interface program and the core program

	Standardization	Defined in each core program
When the container starts	・Trained image recognition model ・How to input a file	・How to load a trained image recognition model
Input during inference	WebAPI input format ・JSON format ・Multipart/form-data format	WebAPI input details ・How to store images in WebAPI input requests, etc.
Output during inference	WebAPI output format ・JSON format	WebAPI output details ・How to store image recognition results using JSON format in WebAPI output

the route over which images are sent to the cloud or data center needs to be separately secured, and steps must be taken to secure this route by without going via the Internet, such as by using a leased line.

One way to address this issue is to perform the image recognition process on a local PC, smartphone, or edge computing device. This prevents images from being transmitted over the network and keeps the images secure because they stay within the device. On the other hand, edge computing devices, PCs and smartphones are insufficiently powerful to perform image recognition, so the recognition process takes a long time. Furthermore, when image recognition is performed on individual devices, even if they do have sufficient performance, it is not possible to consolidate these image recognition processes to increase the efficiency of resource utilization compared to processing performed in the cloud or in a data center.

The DOCOMO Image Recognition Platform solves these issues by setting up a data center (the DOCOMO Open Innovation Cloud) within the NTT DOCOMO communication network as an intermediate between the two, where image recognition processing can be performed. A conceptual illustration of this configuration is shown in **Figure 5**. Since the image recognition process is done in a data center within the NTT DOCOMO communication network, it is possible to communicate with this data center without sending anything via the Internet provided it is accessed via a NTT DOCOMO 4G or 5G line (4th or 5th Generation mobile communication system). This can reduce the risk of image information being leaked to the outside world. On the other hand, since the image recognition functions are centralized at the data center, it is also possible to optimize the computational cost.

In the DOCOMO Image Recognition Platform, connections using this closed network are provided as a Cloud Direct [4] connection option. An additional advantage of this closed network access is that it bypasses the Internet, which means that connections can be made with a shorter delay. The

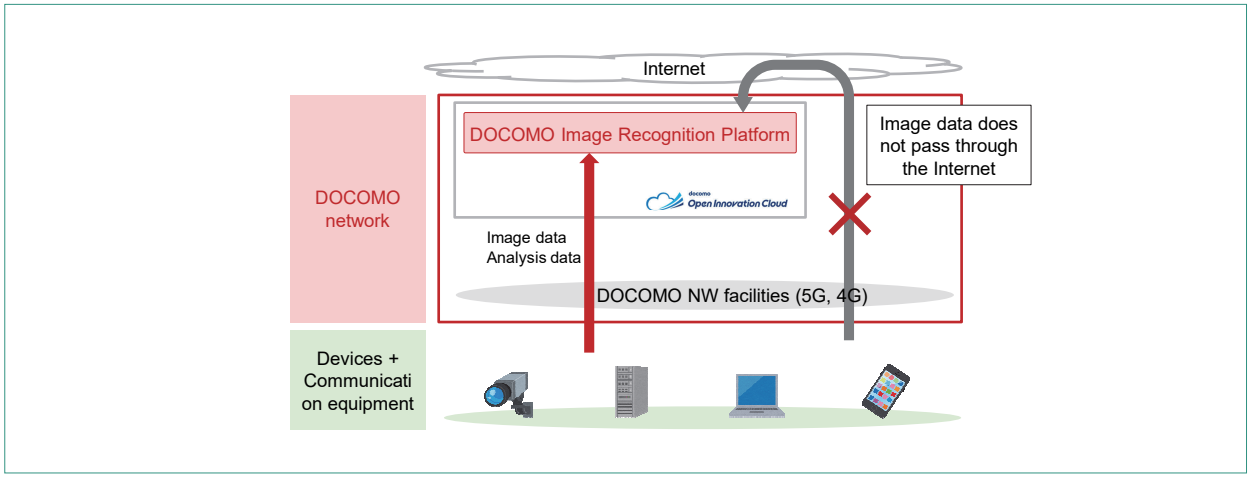


Figure 5 Using closed network connections to make image recognition secure

use of 5G for these connections can reduce the delays still further.

4. Building and Using an Image Recognition API on the DOCOMO Image Recognition Platform

In the DOCOMO Image Recognition Platform, the steps that need to be performed before using the API can be broadly divided into training and deployment. With this service, it is possible to perform these tasks on a web browser (in the cloud) with a simple user interface (**Figure 6**). Training involves creating a model specific to each individual task, and deployment involves using either a trained model or a general-purpose model to perform image recognition. In addition, users can create their own dictionaries for similar image search and specific object recognition functions. A detailed manual for each process is available on the DOCOMO Open Innovation Cloud developer portal.

4.1 Implementing Training

In the current DOCOMO Image Recognition Platform, training can be performed in the object detection and generic object recognition functions. The specific training procedure is described in the tutorial in the developer portal.

In addition to image data, the training process also requires annotation data. This is data that has been tagged with information necessary for the training and evaluation processes of general machine learning[*8], including this image recognition technique. In the training process, the model is

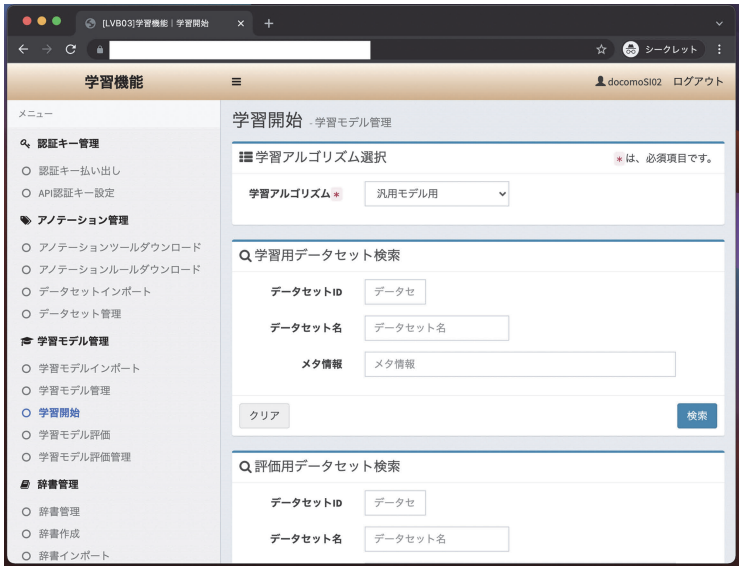


Figure 6 Service admin screen

*8 Machine learning: A technology that enables a computer to learn useful judgment standards through statistical processing from sample data.

trained using annotation data as examples of correct information, and in the evaluation process, the model is evaluated by comparing its inference results with the annotation data. A tool for creating annotation data is provided by NTT DOCOMO as a sample tool.

In the local environment, the user creates a data set consisting of image data and corresponding annotation data according to the specifications and uploads it to the DOCOMO Image Recognition Platform. A verification data set is also created and uploaded in a similar manner, and training is performed by setting up the training data set, verification data set, and target task (object detection or generic object recognition).

Trained models can be evaluated for accuracy on the platform. Like the data sets used for training and verification, an evaluation data set is created and uploaded, and then evaluation is performed by selecting a trained model and an evaluation data set from "Trained model evaluation" in the service admin screen. The evaluation results can be downloaded from the platform. From these results, it is possible to check the indexes showing the accuracy[*9], precision[*10], recall[*11], and F1-score[*12] for both image classification and object detection.

4.2 Implementing Deployment

Deployment makes it possible to use (i.e., perform inference using) models created by training and models independently trained and provided by NTT DOCOMO via a WebAPI. A deployed image recognition AI can be used by dispensing and assigning an API key. A REpresentational State Transfer (REST) API[*13] is adopted as the API method and can be linked with the user-side system with a simple design.

A deployed image recognition AI can be managed by adjusting the API management and API authentication key settings in the service admin screen. In API management, it is possible to manage trained models and add or delete image recognition AI. These functions can be easily scaled. In API authentication key setting, it is possible to assign and change the authentication key of a deployed image recognition AI.

4.3 Implementing Dictionary Creation

On the DOCOMO Image Recognition Platform, it is possible to create unique dictionaries for similar image search and specific object recognition. In similar image search, it is possible to determine which images in a pre-prepared dictionary of images resemble the requested image. Furthermore, specific object recognition can identify what is depicted in the requested image by comparing it with images in a pre-prepared dictionary.

The preparation of both types of dictionary involves collecting and annotating images. After a dictionary has been created, each function can be used by deploying it in the same way as other image recognition functions.

5. Conclusion

In this article, we described the background and challenges of making image recognition services easy to use, and we discussed the image recognition functions provided by the DOCOMO Image Recognition Platform. We described the character-

*9 Accuracy: The percentage of inferred data that is correctly classified and detected.
*10 Precision: In object detection, the percentage of detected objects that are certain to be objects corresponding to a given label. In generic object recognition, the percentage of data inferred to correspond to a certain label for which these inferred results are correct.
*11 Recall: The percentage of all data for a label that is correctly classified with that label (in object detection).
*12 F1-score: The harmonic mean of precision and recall.
*13 REST API: An API that adheres to the REST style of software architecture, which evolved from design principles proposed by Roy Fielding in 2000.

istics of the system that enable the provision of various image recognition functions in the DOCOMO Image Recognition Platform, and way in which it improves the security of image recognition by performing recognition processing in a closed network. We also showed how the DOCOMO Image Recognition Platform is actually used in practice. Going forward, NTT DOCOMO will continue to improve and update its functions to provide the image recognition functions that users need.

REFERENCES

[1] AWS: "Amazon Rekognition Custom Labels." https://aws.amazon.com/rekognition/custom-labels-features/?nc1=h_ls
[2] Google Cloud: "AutoML Vision." https://cloud.google.com/vision/automl/docs
[3] NTT DOCOMO: "The DOCOMO Image Recognition Platform." https://www.nttdocomo.co.jp/biz/service/dirp/
[4] NTT DOCOMO: "What is Cloud Direct?" https://developer.dev-portal.d-oic.com/document/docs/cloud-direct/concepts/overview.html

Technology Reports (Special Articles)
Tourist Spot Recommendation System | Tourism Informatics | Neural Network
Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

"Generic POI Recommendation": A Brand-new Deep Learning Approach for Discovering Potential Sightseeing Spots

X-Tech Development Department Hisao Katsumi Wataru Yamada Keiichi Ochiai

Data mining each geographic region for new tourist spots using AI technology can serve as a new solution to social issues such as the overconcentration of tourists in certain spots and having difficulty in attracting tourists to regional cities. In this article, we propose a new idea called generic POIs, alternative sightseeing spots to famous spots. In addition, we propose a method to discover POIs that look similar to those of existing famous sightseeing spots as generic POIs. We also present results of actual mining of Web data for generic POIs and explain our evaluation of the results. The efforts introduced in this article hold great promise as a new step toward realizing a tourism industry that utilizes AI technology going forward.

1. Introduction

The rapid development of means of transportation and the proliferation of mass media have made it possible for us to travel to a variety of places for sightseeing. However, this has led to the issue of overtourism, the overconcentration of tourists in certain tourist sites [1]. Overtourism is a serious issue that adversely affects both tourists and residents living in those sightseeing areas. Three methods are being carried out to effectively address this issue: (1) increase the tourism carrying

capacity of existing tourist spots, (2) deconcentrate tourists spatially and temporally at existing tourist spots, and (3) create new tourist spots and induce tourists and visit those destinations [2].

Various efforts are being carried out in the area of (1) increasing tourism carrying capacity, such as building parking facilities at tourist spots. For (2), typical examples of efforts include methods to prevent congestion by presenting real-time crowdedness information in online recommendations of tourist spots [3] and methods to mine for spots that are "well-kept secrets"—sites preferred by local residents more than tourists—by analyzing the attributes of users who post images of tourist attractions on photo sharing websites [4]. Efforts to create new tourist spots and induce tourists and visit them (effort 3) are also important. These efforts include increasing the maximum number of tourist spots in a city by considering spots that had not considered to be tourist sites as tourism resources locations.

As an effort in (3) to create new tourist spots, we at NTT DOCOMO focused on the phenomenon of spots' drawing buzz because they resemble already famous spots and becoming even more popular as a result. For example, Takeda Castle Ruins, famous in Japan, drew media attention because it appears similar to Machu Picchu. As a result, the number of tourists to that spot grew. In this way, if we can find tourism value in spots that had not been considered to date as tourist attractions from the standpoint of their resemblance to famous tourist spots, then it is possible to mine for new tourist spots. At NTT DOCOMO, we call these spots, which can serve as substitutes for already famous spots due to their resemblance, "generic POIs."

In this article, we describe techniques of using image processing AI technology and other technologies to present spots resembling already famous tourist spots as generic POIs. Furthermore, we explain evaluation of our proposed method using images of spots on the Web.

2. Generic POI Extraction Technique

Figure 1 provides an overview of our proposed method. It collects images of candidate spots for mining generic POIs and images of existing famous tourist spots from the Web and calculates the degree of similarity in all combinations of images. Candidate spots that are included in combinations ranked high in similarity are output as generic POIs.

This method and the idea behind it seek to mine for new tourist spots using only images of spots. Compared with general methods for mining tourist spots, our method can mine new tourist spots without depending on the amount of data collected in advance such as word-of-mouth comments and posted images. Furthermore, the method makes it possible to discover new tourism value in spots that even local residents had not recognized as tourist attraction.

3. Generic POI Extraction Technique

This method is composed of the following four processes (Fig. 1).

Process (1): Collect from the Web images of

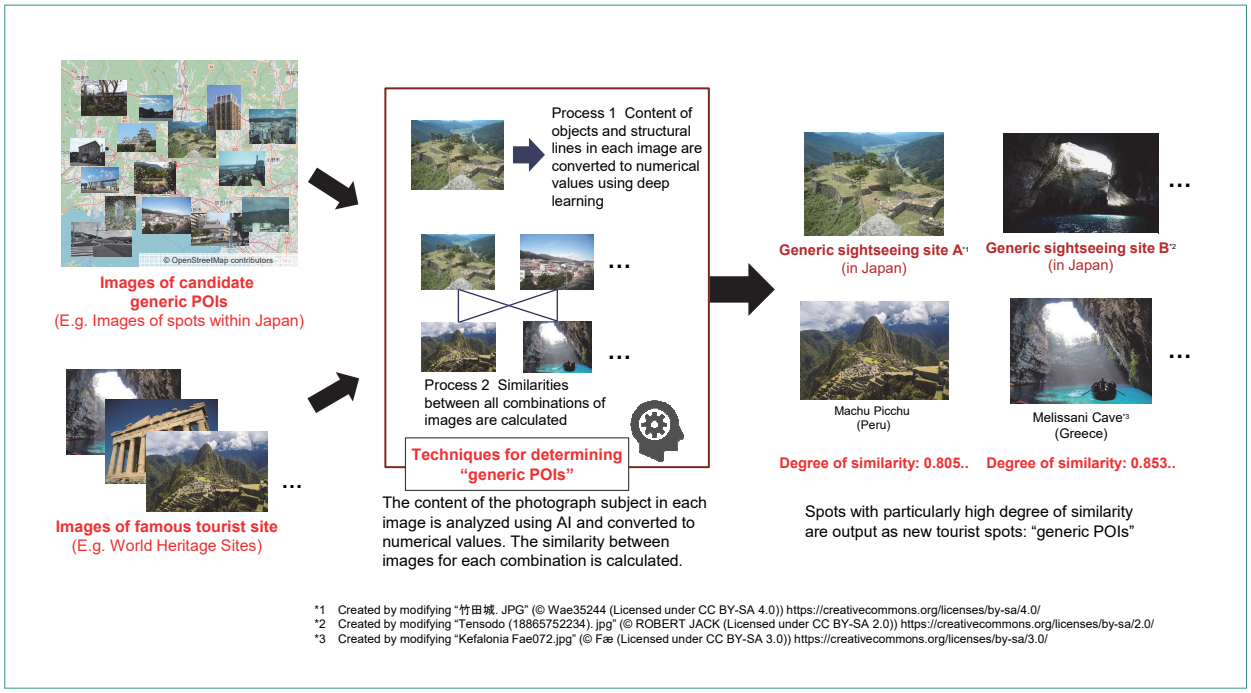


Figure 1 Overview of proposed method

candidate spots taken within an arbitrary distance range.

Process (2): Collect from the Web images of famous spots based on their names.

Process (3): Calculate the similarities between collected images of candidate spots and famous spots for all combinations.

Process (4): Output candidate spots included in the top N combination of similar images as generic POIs.

We assume the use of a variety of images of spots obtained from the Web for Processes (1) and (2). For evaluation of our method, we used images of tourist spots posted on the photo sharing site Flickr[*1].

For Process (3), we calculated image similarity for all combinations of candidate spot images and famous spot images. To calculate image similarity, we converted each image of candidate spots and famous spots into feature vectors[*2] and used cosine similarity[*3] between two feature vectors as the measure of similarity. For feature vectors of tourist spot images, techniques to extract semantic features such as content describing an image and techniques to extract structural features such as structural lines are considered to be useful. For the former, Visual concept is a technique that quantifies the extent to which each image contains each component of scenery such as mountains and rivers in 365 categories [5]. For the latter, computing GIST descriptor is a method to extract feature values such as structural lines and distribution of light and dark areas from an image [6].

Finally, in Process (4), we output candidate site images in top-ranking combinations of images from

*1 Flickr: Trademark of Oath Inc. in the U.S.

*2 Feature vector: Representation of patterns and features in data as a vector, which is an array of numerical values, allowing the data features to be handled programmatically.

*3 Cosine similarity: Numerical measure of how close the directions of two vectors are.

all combinations as generic POIs. This method considers as generic POIs to be candidate spots in the top N ranked combinations in image similarity.

4. Evaluation Experiment

To study the validity of our proposed method, we created evaluation data. Furthermore, we mined the evaluation data for generic POIs using the proposed method and evaluated the results by calculating error as the distance between the location of a candidate site and the correct example.

4.1 Creation of Evaluation Data

An overview of the method of creating evaluation data is shown in **Figure 2**. First, five famous tourist spots in Japan already known for their resemblance to already famous overseas spots (see table on the left in Fig. 2) were designated as correct examples of generic POIs.

Next using these correct examples, candidate spot images within geographic range of the correct examples of generic POIs were collected from Flickr. Specifically, images taken within a standard region mesh*4 belonging to each of the five correct examples were collected as candidate spot images. For the standard region mesh, we used a secondary mesh that divides the geography of Japan into 10 × 10 km squares. As shown in the right of Fig. 2, the location of a correct generic sightseeing area is marked by ★. Spots where images of candidate spot images were taken were plotted and marked by ●. Using this method, 2,353 candidate spot images were obtained.

To obtain images of the famous spots belonging to the five correct examples, for each of the five spots we entered its English name into the query*5 box in Flickr to get five images each, for a total of 25 images of famous spots. However, the images obtained from search results of famous spots

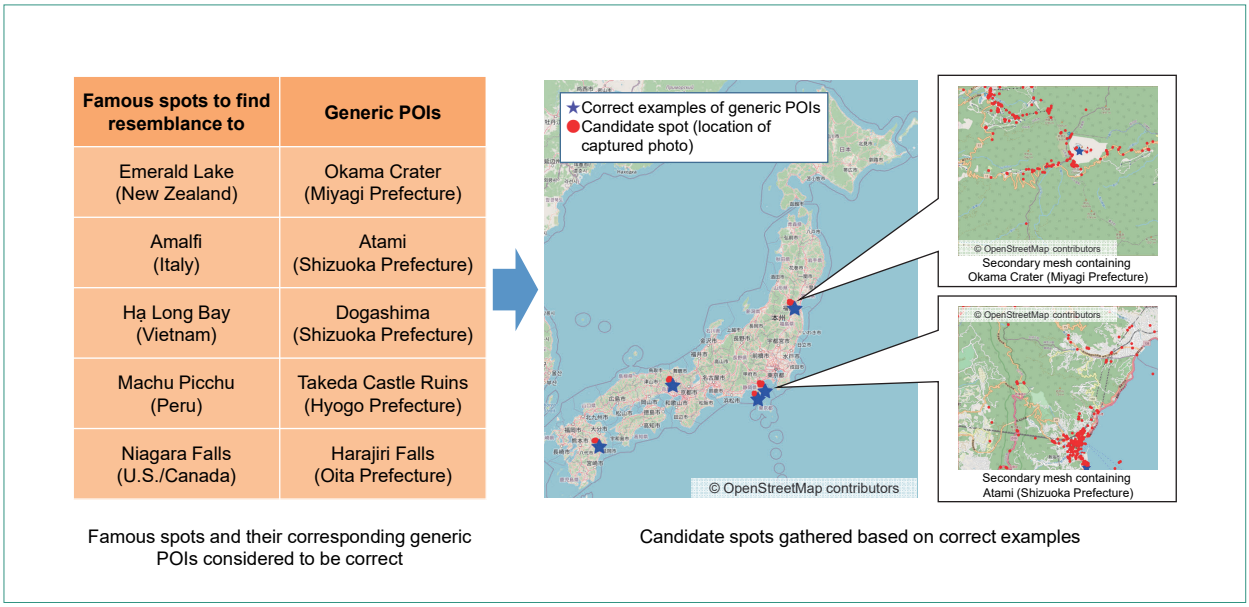


Figure 2 Creation of evaluation data

*4 Standard region mesh: Geographic division of Japan into a mesh according to latitudes and longitudes for the purpose of aggregating data in each region.

*5 Query: A database query (processing request).

included not only images of the spots themselves but also extraneous images such as people and food, which are considered noise. These noise images were removed manually to prepare the spot images as evaluation data.

Please note that the images collected as described above to create evaluation data were only those licensed by their owners to allow their use for research purposes. Images on Flickr that had the appropriate Creative Commons license*6 type were collected.

4.2 Feature Vectorization of Spot Images

To mine for generic POIs proposed in this article, feature vectorization is carried out for each spot image. The degree of similarity between images is then calculated. For evaluation, we used the following three techniques for feature vectorization and compared the output results obtained.

Technique (1): Visual concepts of scenes

In this method, a neural network*7 is used to determine the extent to which content in each spot image belongs to the 365 categories of scenery, such as mountains, rivers, and the sea. Each spot image is converted into 365-dimensional feature vector. The neural network used here is ResNet18. It carried out training using the Places365 scene recognition dataset [7].

Technique (2): GIST descriptors

Each spot image was converted to a feature vector using GIST. GIST features are extractions of structural features in a spot image, such as the rough distribution of light and dark areas.

Technique (3): Embedding vectors*8 of scene categories

Considering that a neural network extracts features in stages, we used a 512-dimensional vector as the feature vector at the stage before the neural network used in Technique (1) finally computes a 365-feature vector as the visual concept. The feature vector obtained in this manner has higher-level abstraction embedded just before the neural network makes the final inference of 365 scene categories as visual concept (Technique (1)). This processing makes it easier to consider as similar structures that would be classified into different categories under the 365 scene categories.

4.3 Evaluation Results and Discussion

From evaluation data created using the above methods, the proposed method mined for N (N = 10, 20, 30) generic POIs. Photos of generic POIs mined using Technique (3) with N = 30 are shown in **Figure 3** and their locations are shown in **Figure 4**. Within each mesh, we calculated the distance (km) of each mined generic sightseeing site from location of the correct example and obtained an average distance, which we consider as a metric of error. As a generic sightseeing site, the closer in location the mined candidate site is to the correct example, the smaller we consider the error to be. When a spot completely matching the correct example in location is mined, the error becomes 0. However, because spot images collected from the Web includes images taken from different perspectives, images of sites mined from the correct example that are in actuality images of the same

*6 Creative Commons license: Presents creator's intent regarding use of their copyrighted work in accordance with rules set forth by Creative Commons.

*7 Neural network: Mathematical model that imitates the workings of the biological brain, allowing recognition of numerical patterns and making inferences.

*8 Embedding vector: A vector where necessary information is converted from higher-dimensional features to lower-dimensional features.

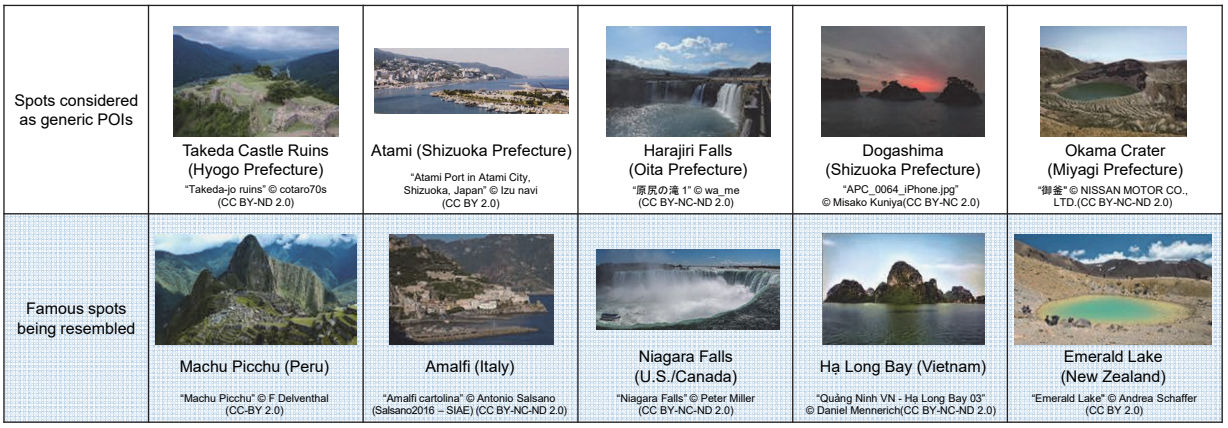


Figure 3 Examples of spots mined as generic POIs using Technique 3 ($N = 30$)

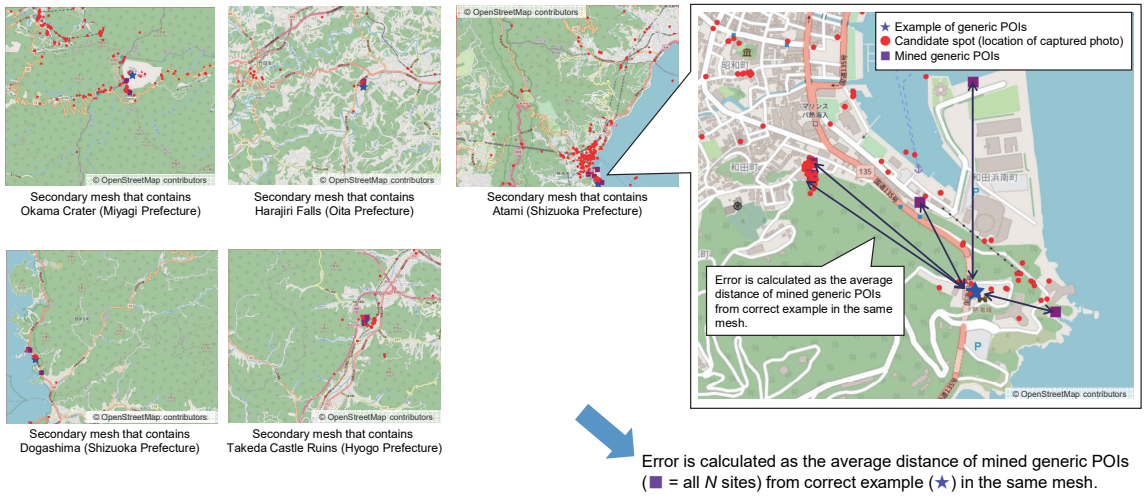


Figure 4 Calculation of error in location of spots mined as generic POIs using Technique 3 ($N = 30$)

site may not necessarily have an error of 0. As shown in the right side of Fig. 4, mining the Atami area resulted in several images of spots of Atami that were images of the correct example, a site that resembled Amalfi. These results indicate that these spots were mined as generic POIs of Amalfi. The final results of error evaluation for all the combinations of feature vectorization technique and N candidate images are shown in **Table 1**.

From the results shown in Figures 3 and 4 and Table 1, we see that candidate spots that appear roughly similar to famous spots can be mined as

Table 1 Error calculated for each combination of feature vectorization technique and *N*

	Average error in distance for top 10 spots (km)	Average error in distance for top 20 spots (km)	Average error in distance for top 30 spots (km)
Technique (1) Visual concept	2.18	2.20	2.22
Technique (2) GIST	2.56	2.35	2.45
Technique (3) Embedding vector	0.33	0.27	0.34

generic POIs. As shown in Table 1, Technique (3) produced results that exceeded those of the other techniques. Furthermore, comparing Techniques (1) and (2), in all cases Technique (1) produced better results than Technique (2). These results suggest semantic features, such as content about the subject of a spot photo, are more effective than structural features such as structural lines in an image.

Because our method does not carry out processing of candidate spot images such as clustering*9, it can be applied regardless of the content of input candidate spot images and famous spot images and their quantity. Going forward, we plan to evaluate whether similar results can be obtained when the scale of evaluation data is increased.

5. Conclusion

In this article, we described the idea of data mining for new tourist spots, which we call generic POIs for their resemblance to existing famous spots. We also proposed mining techniques using a neural network, and described evaluation of applying these techniques to tourist spot images posted on the Web. However, as our proposed method was implemented based on five correct locations collected on the Web, we plan to study whether our method can correctly mine a large amount of candidate spot images and famous spot images to find generic POIs, as well as whether mined generic POIs can improve the issue of the overconcentration of tourists and induce them to visit the newly discovered areas.

The proposed ideas and techniques described in this article hold great promise as the first step of efforts that use AI technology to effectively discover new tourism value in spots that had not been considered tourist attractions. They can be expected to be useful for not only the issue of overtourism but also for regional revitalization [8]. In addition, we hope that our proposed ideas and techniques can serve as a step in establishing the new normal*10 in tourism lifestyle, as people accept a new lifestyle that is conscious of preventing the spread of COVID-19. In this new tourism lifestyle, instead of traveling to distant overseas and domestic locations, tourists will see a variety of local spots with a new perspective, becoming cognizant of their heretofore unrealized tourism value and appreciating their hidden allure.

REFERENCES

[1] M. Duignan: "'Overtourism'? Understanding and Managing Urban Tourism Growth beyond Perceptions: Case

*9 Clustering: Division of a large amount of data into data groups with similar characteristics.

*10 New normal: A state in which a new common sense has irreversibly taken hold as a result of changes in the social environment and circumstances.

Studies,” United Nations World Tourism Organisation (UNWTO), pp.34–39, Mar. 2019.

[2] T. Mainil, E. Eijgelaar, J. Klijs, J. Nawijn and P. Peeters: “Research for TRAN Committee-Health tourism in the EU: a general investigation,” European Parliament, Directorate General for Internal Policies, 2017.

[3] M. Hidaka, Y. Kanaya, S. Kawanaka, Y. Matsuda, Y. Nakamura, H. Suwa, M. Fujimoto, Y. Arakawa and K. Yasumoto: “n-site Trip Planning Support System Based on Dynamic Information on Tourism Spots,” Smart Cities, Vol.3, No.2, pp.212–231, Apr. 2020.

[4] C. Zhuang, Q. Ma, X. Liang and M. Yoshikawa: Anaba An obscure sightseeing spots discovering system,” 2014 IEEE International Conference on Multimedia and Expo (ICME), pp.1–6, Sep. 2014.

[5] C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Penas and V. Petras: “Evaluating Systems for Multilingual and Multimodal Information Access,” 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, Revised Selected Papers, p.527, 2008.

[6] A. Oliva and A. Torralba: “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope,” International journal of computer vision, Vol.42, No.3, pp.145–175, May 2001.

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba: “Places: A 10 Million Image Database for Scene Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.40, Issue 6, pp.1452–1464, Jul. 2017.

[8] NTT DOCOMO press release: “The first time in Japan! Photo contest to discover scenery in Shikoku that look just like famous overseas spot. In a time when it is difficult to travel far, discover places that give you a sense of traveling abroad! AI judges the similarity of scenery images,” Jul. 2021 (In Japanese).

Technology Reports (Special Articles) 5G Edge AI Self-driving

Special Articles on AI—Expansion of AI Technologies to Diverse Industries and Basic Technologies Supporting AI Applications—

Self-driving Support System toward Personal Mobility Using Edge AI-compatible 5G Device

Communication Device Development Department Makoto Takahashi Yuuki Nakazawa Tomoya Moribe Kenya Ikeda

Expectations are growing for diverse self-driving services using personal mobility vehicles and service robots with the hope of solving a variety of social problems brought on by an aging society and labor shortages. Ensuring safety will be of prime importance in implementing these services in society—obstacle-avoidance performance needs to be improved and remote-steering services need to be supported. In response to these needs, NTT DOCOMO has developed technologies for obstacle detection on an edge device with high real-time performance and for secure and low-latency remote steering using 5G communications in a closed environment. These technologies support safe self-driving thereby contributing to the implementation of self-driving services in society.

1. Introduction

The total population of Japan as of October 1, 2020 was 125,710,000. As part of this figure, the population of those 65 years of age or over came to 36,190,000 bringing the percentage of elderly people in Japan to 28.8%. Statistical data from the Cabinet Office, Government of Japan indicates that approximately one in 2.6 people will be 65 years of age or over by 2065 [1]. In response to the aging society, there are expectations in Japan for the social implementation*1 of labor-saving services such as self-driving wheelchairs enabling the elderly to move about freely and self-driving robots for making

last-mile deliveries to one's home. Ensuring safety in the social implementation of self-driving services is of prime importance, so initiatives for improving safety are needed. In addition to distance sensors such as conventional Light Detection And Ranging (LiDAR)*2, these will include an obstacle-detection function capable of instantaneous response using image recognition technology*3 operating on an edge device and support for a remote-steering mode to enable self-driving in situations where autonomous driving*4 is difficult.

With a view to safe self-driving in various types of personal mobility*5 vehicles and service robots, this article describes a self-driving support system toward personal mobility achieved by using a remote-steering system constructed in NTT DOCOMO's Multi-access Edge Computing (MEC)*6 environment in combination with an Edge AI*7-compatible 5G device.

2. System Overview

The overall configuration of a self-driving support system developed by NTT DOCOMO toward personal mobility is shown in **Figure 1**. This system consists of three functional blocks: (1) "Edge AI-compatible 5G device" as an add-on to various types of personal mobility vehicles and service robots, (2) "Edge AI applications" for performing processing requiring real-time performance on that device, and (3) "remote steering system" that supports the monitoring and driving of personal-mobility self-driving vehicles. The system takes on a configuration that can be used not for the development of personal mobility itself such as wheelchairs but rather as an add-on to vehicles developed by partner companies. It enables self-driving support that enhances safety for diverse types of personal mobility vehicles and service robots.

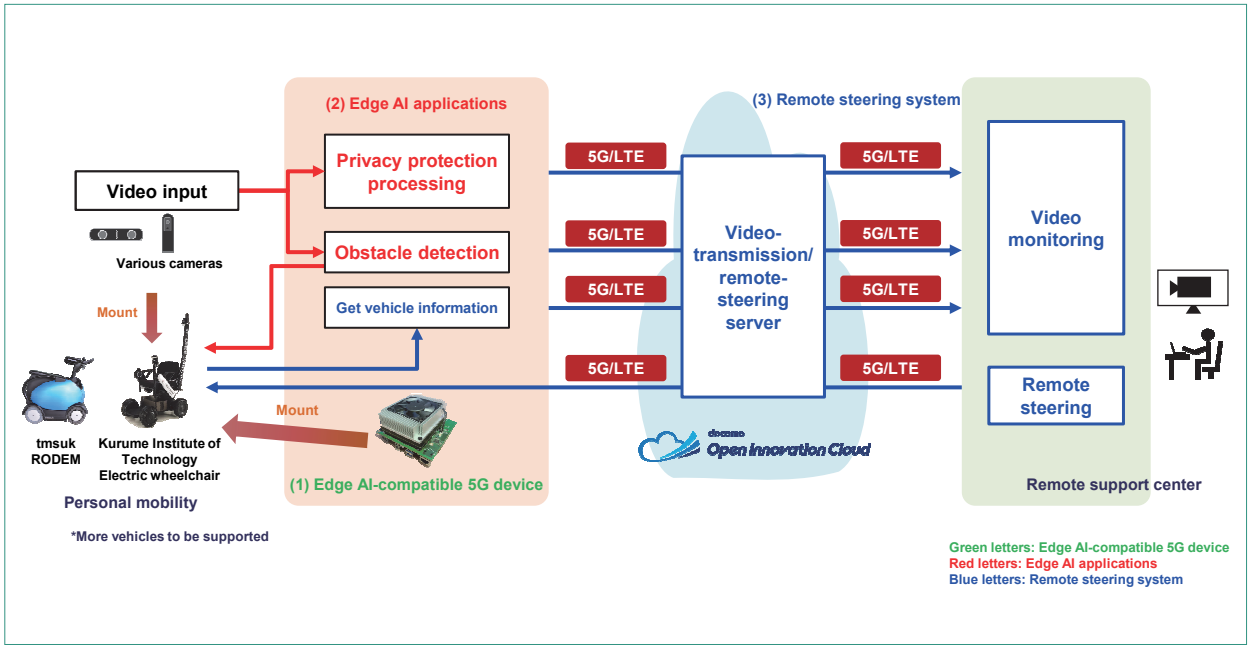


Figure 1 Overall configuration of a "self-driving support system toward personal mobility"

*1 Social implementation: In this article, the state in which a mechanism for solving social issues penetrates social life and continues to be used.

*2 LiDAR: Sensor technology that irradiates objects in the surrounding area with light such as near-infrared light and enables object detection by reflected light.

*3 Image recognition technology: Technology for mechanically understanding images and extracting meaning using image processing technology, machine learning technology, etc.

Specifically, this system will enable appropriate vehicle control such as by recognizing that a certain obstacle in the vicinity of a personal mobility vehicle is a person to be avoided. It can also enhance the safety of a self-driving system that uses only distance sensors such as LiDAR. Furthermore, with a view to the social implementation of self-driving services, there will be a need for a function that can remotely support self-driving to improve the safety of autonomous driving technology in the field of personal mobility and to deal with unforeseen problems during autonomous driving. This system transmits on-site video to a remote support center, which is equipped with a function for monitoring vehicle conditions. In this way, it becomes possible for a manager at the remote support center to perform remote steering in various situations, such as when the personal mobility vehicle loses self-position during autonomous driving due to a crowd of people or obstacles in the periphery or when the vehicle moves into a location in which no digital map exists during autonomous driving. This system will also consider the need for privacy in the case of people captured by on-site video, so it will enable the faces of any persons appearing in the video transmitted to the remote support center to be preprocessed by mosaic masking.

The cloud environment used for constructing this system is the DOCOMO Open Innovation Cloud® [2], an MEC environment provided by NTT DOCOMO. A direct connection to this cloud can be made using the 5th Generation mobile communications system (5G), which enables high-definition and low-latency video transmission unique to 5G as well as high security by not passing through the Internet.

3. Edge AI-Compatible 5G Device

3.1 Device Overview

This device consists of a high-performance processor, 5G module, and carrier board*8 equipped with various interfaces as described below. On applying this device to personal mobility, it must be capable of high-load processing as in the case of image recognition on an edge device and of large-capacity, low-latency communications for transmitting high-definition video. It must also be relatively easy to incorporate in the vehicle. By combining the constituent elements described below, this device can be used to support self-driving in personal mobility.

3.2 High-performance Processor Capable of Edge AI Processing

Autonomous driving in personal mobility requires multiple types of AI processing on an edge device excelling in real-time performance. While AI processing such as image recognition is generally executed on the cloud due to its superior machine performance, this device mounts the Jetson AGX Xavier [3] processor for edge computing*9 from NVIDIA enabling AI processing on an edge device. It also supports mounting of the Jetson Xavier NX [4] processor featuring lower power consumption and a lower price. The user can therefore choose which of these two processors fits the target application.

3.3 5G Module for High-speed, Large-capacity, Low-latency Communications

We adopted the Telit FN980 [5] module supporting 5G communications for which interoperability

*4 Autonomous driving: A function for driving to a destination while estimating self-location using a digital map and distance sensors, motion sensors, etc. mounted on a personal mobility (see *5) vehicle.
*5 Personal mobility: In this article, an electrically driven vehicle for personal use.
*6 MEC: A mechanism of installing servers or storages within a carrier network, at locations near users.
*7 Edge AI: In this article, AI processing not only in an MEC environment but also on actual devices.
*8 Carrier board: Hardware having essential input/output interfaces.

tests with the NTT DOCOMO network have been completed. In the area of self-driving support for personal mobility through remote steering, measures have been taken to improve safety by using multiple cameras to provide high-definition video, 360-degree video, etc., but the deployment of 5G has made the transmission of such video all the more feasible.

This module also supports an external antenna so that stable communications can be achieved through an antenna external to a personal mobility vehicle while embedding this device in the interior of the vehicle. The module also adopts a general-purpose M.2 connector*10 as a communication-module interface to ensure connectivity with various types of equipment.

3.4 Carrier Board Equipped with an Interface for External Equipment Connections

We developed a carrier board to integrate the high-performance processor and 5G module described above. This board incorporates abundant interfaces such as a serial communications interface and High-Definition Multimedia Interface (HDMI) input (Jetson AGX Xavier only) as listed in **Table 1**. The idea here is to enable connections to smartphones and diverse types of sensors including external cameras, which is difficult for ordinary routers. Moreover, envisioning the use of this device as an add-on to personal mobility vehicles, it has been given a form that emphasizes ease of embedding. This device has general-purpose features that support embedding in various types of vehicles and not just specific types of personal mobility.

4. Edge AI Applications

4.1 Overview of Applications

With the aim of improving the safety of self-driving

Table 1 Specifications of carrier board

Size	120 × 120 mm
Operating temperature	-20 – +80℃
Operating humidity	10 – 90%
Interfaces	· 1×HDMI Type A (output) · 1×USB3.1 Type C · 2×USB3.1 Type A · 1×USB2.0 Micro B (OTG) · 1×RJ-45 for GbE · 1×Micro SD slot · 1×DC-in 9 – 19V (6 pin Euroblock) · 1×M.2 B-Key 3052 (Telit FN980m 5G/LTE) · 1×M.2 M-Key 2280 (SC710N1 M2 HDMI) (1×HDMI Type A input) · 1×M.2 M-Key 2280 (enables addition of M.2 external storage) · 1×nano SIM slot
Extension pin headers	· 1×CANBUS · 1×UART 3.3V/5V TTL (JST-GH 6 pin) · 1×UART 3.3V/5V TTL (DF-13-6 pin) · 1×RS-232/1×I2C/5×GPIO · 1×MIPI CSI connector (120 pin) · 1×front panel (reset/recovery/power ON)

*9 Edge computing: Technology that distributes edge servers closer to the users to improve response and reduce latency.
*10 M.2 connector: A connection terminal applicable to thin-type, high-performance devices.

in personal mobility, NTT DOCOMO has developed applications to be run on an edge device. These include peripheral video recognition to avoid collisions with obstacles, which requires real-time performance, and privacy protection processing with respect to persons caught in video transmitted for remote-steering purposes.

4.2 Detection of Obstacles in Vicinity of Personal Mobility Vehicle

In autonomous driving, there is a need to accurately determine the presence of any obstacles in the vicinity of the personal mobility vehicle. To this end, the system performs image recognition and analysis using stereo-camera video to identify the type of obstacle and calculate the distance to it. Furthermore, if the obstacle happens to be a moving object such as a human being, the system predicts the direction and amount of movement. The information output from these various forms of video analysis will be instantaneously passed from the edge device to the personal mobility vehicle without any cloud intervention thereby enabling fast vehicle control and safer self-driving.

To improve the operability of a remote operator and to improve system safety in remote steering, the system performs 360-degree image recognition and analysis about the personal mobility vehicle and displays the results on a screen at a remote site. In this way, the remote operator can visually ascertain whether someone is in the vicinity, and if so, at what position.

4.3 Protection of Private Information

When transmitting video from a medical facility, for example, there are cases in which privacy protection is required with respect to any persons that happen to be included in the transmitted video. To deal with such a situation in this system, we made it possible to perform mask processing on the edge device itself with respect to any persons detected in the video before transmitting that video to the remote operator. Here, considering that total masking of a person may impair the remote operator's visibility, the system performs mask processing on only the face of a detected person. However, when a person's face itself becomes the target of image recognition and processing, there is a tendency for the accuracy of privacy protection to drop since the target area is relatively small. For this reason, the system uses person detection and face detection in a stepwise manner.

4.4 Reducing Latency in Transmitting Video and Control Information

In autonomous driving and remote steering of a personal mobility vehicle in which collisions must be prevented and safety ensured, it is vitally important to achieve low latency in addition to dealing with high-load processing. Additionally, given that multiple AI processes are being performed in parallel on an edge device, the limited computational resources on that device must be efficiently used. Consequently, to achieve low-cost and speedy development of this system, we used Yolo [6], an open-source AI framework*11, and devised measures to accelerate the processing of Edge AI applications.

In inference processing, this system forgoes single-precision floating-point operations*12 and adopts half-precision floating-point operations*13 instead. For example, in the case of a recognition target

*11 Open-source AI framework: A compilation of diverse programs to simplify AI development work is called an AI framework. At present, a variety of AI frameworks can be used without charge as open source software.

*12 Single-precision floating-point operations: 32-bit floating-point operations abbreviated as FP 32.

*13 Half-precision floating-point operations: 16-bit floating-point operations abbreviated as FP 16.

having a size comparable to that of a human being, it is known that the use of single-precision floating-point operations, though improving computational accuracy, does not greatly improve inference accuracy. We therefore decided to use half-precision floating-point operations since they can perform inference processing faster than single-precision floating-point operations. We also optimized each processing parameter in Edge AI applications, optimized the buffer size for decoding, and performed downsampling*14 of input images. In these ways, we made more efficient use of all computational resources running on the edge device and achieved low-latency operations.

5. 5G Remote Steering System

5.1 System Overview

This system consists of an Edge AI-compatible 5G device (Fig. 1 (1)), video-transmission/remote-steering server constructed on the DOCOMO Open Innovation Cloud, web/iPad*15 application for monitoring, and a controller for steering (Fig. 1 (3)). In video transmission, the edge device transmits the video and various types of sensor information created by an Edge AI application to the remote support center via the server for display on the web/iPad application for monitoring purposes. In remote steering, the steering information input by the controller is transmitted via the server to the edge device, which sends it on to the personal mobility vehicle. Here, communications are performed via the 5G/LTE network using a direct connection to the server on the DOCOMO Open Innovation Cloud. Issues surrounding the practical implementation of remote steering as uncovered during the study of this system are listed in Table 2 and discussed below.

5.2 Low Latency and High Security Using the DOCOMO Open Innovation Cloud

From the viewpoint of low latency, operations from the input of video to the input/reflection of steering information must be completed within 500 ms so that the operator on the remote-support-center side can perform remote steering safely. We satisfied this requirement for low latency within 500 ms by devising measures for processing on the edge device as described above as well as by combining the DOCOMO Open Innovation Cloud, NTT DOCOMO's original MEC platform, with a direct cloud connection over a dedicated line. Constructing this remote steering system on the DOCOMO Open Innovation Cloud in this way makes for improved low-latency characteristics compared with the use of a public cloud*16 on the

Table 2 Issues in achieving a practical remote steering system

(1) Low latency	Low latency from checking conditions to reflecting steering is essential.
(2) 360-degree safety check	Visual recognition of people or objects approaching from the side or rear is essential.
(3) Vehicle information check	Monitoring of vehicle information such as remaining battery capacity is essential.
(4) Flexible controller support tailored to vehicle	Supporting a specific controller tailored to the vehicle is essential to enhancing system extensibility.

*14 Downsampling: The process of lowering the resolution of an image to reduce computational complexity.

*15 iPad: Apple, the Apple logo and iPad are trademarks of Apple Inc. registered in the United States and other countries. TM and ©2020 Apple Inc. All rights reserved.

*16 Public cloud: Cloud computing services that anyone can use over the Internet.

Internet. In trials conducted with vendors, we found that the time taken from transmitting video by an Edge AI application including privacy protection processing to its display at a remote site is about 300 ms while the time taken to transmit steering information is about 100 ms for a total of about 400 ms, which confirms that safe remote steering could be achieved within 500 ms. Another advantage of this system in addition to low latency is that the communication path has a closed configuration not open to the NTT DOCOMO network. The system is therefore separated from the outside, which reduces the risk of unauthorized access thereby achieving high-security characteristics.

5.3 Visualization of Video and Sensor Information

Next, from the viewpoint of ensuring safety during remote steering, it is important to not only secure a field of view in the forward direction but to also visually recognize persons and objects approaching from the side or the rear. Additionally, information on the condition and orientation of the personal mobility vehicle obtained from battery information, sensors, etc. must be monitored to ensure safe driving. These problems are addressed by displaying both webcam video in the forward direction and omnidirectional video from a 360-degree camera on the web/iPad application for monitoring. There is also a function for sending arbitrary text and numerical information from the edge device to the server for visualization on a web application (**Figure 2**). As for the 360-degree camera video, the viewpoint can be controlled by a drag operation so that the remote operator can zoom in on a particular section of the surrounding environment. Also provided is a radar-display function for presenting positional information on people detected by an AI application from 360-degree camera video. Additionally, with regard to the above

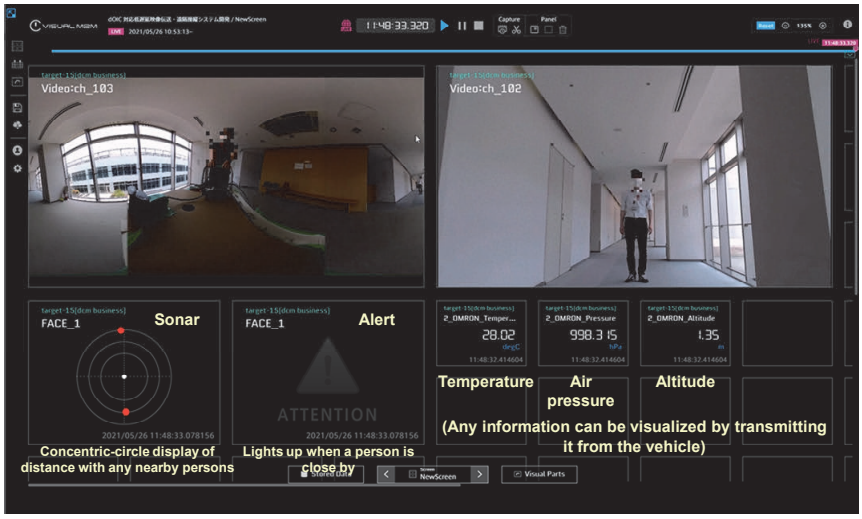


Figure 2 Screenshot of video-transmission/remote steering system

function for visualizing arbitrary data, the control application of each personal mobility vendor may transmit any information targeted for visualization by using a specific Application Programming Interface (API)*17. Each vendor may freely select the information to be visualized, the types of sensors to be used, etc.

5.4 Flexible Support of Multiple Mobility Controllers

Next, from the viewpoint of system operability, it must be possible to use diverse types of controllers specific to various types of personal mobility vehicles. In response to this problem, the specifications of this system allow for transparent transmission of steering information to the server instead of having to convert that information to some type of unique format. Here, the system transmits steering information with low latency via the DOCOMO Open Innovation Cloud so that remote steering can be achieved by incorporating that steering information in the control application developed by each personal mobility vendor. The vendor only needs to select the controller equipment conforming to the control application, which enables the system to provide flexible support for controllers applicable to each personal mobility vehicle.

6. Conclusion

With the aim of solving social problems brought on by an aging society and labor shortages, this article described a self-driving support system toward personal mobility that will contribute to safe self-driving of personal mobility vehicles and service robots.

It described, in particular, the features of an "Edge AI-compatible 5G device" that connects to a webcam and 360-degree camera and achieves image recognition processing and low-latency transmission of high-definition video for remote monitoring through 5G communications, "Edge AI applications" that enable obstacle detection for real-time vehicle control and privacy-protection processing against transmitted video to be performed on an edge device, and a "remote steering system" using the DOCOMO Open Innovation Cloud that enables secure and low-latency remote monitoring and steering.

A variety of trials have been held in this area and some commercial services on a chargeable basis have begun to be provided. However, the cost incurred by a total system including hardware and software is still high, so to make further inroads in society, the service-provision price must be reduced and more improvements must be made in system safety. Going forward, it will become increasingly important to reduce costs by developing a general-purpose system that incorporates low-cost and commercially competitive and available equipment and technologies in addition to NTT DOCOMO's 5G network and MEC platform. NTT DOCOMO will promote the creation of new value and the social implementation of new technologies in collaboration with co-creation partners toward the solution of diverse social problems.

REFERENCES

[1] Cabinet Office, Government of Japan: "2021 White Paper on the Aging Society (Summary), Section 1 Situation of the Aging Population," (In Japanese). https://www8.cao.go.jp/kourei/whitepaper/w-2021/

*17 API: An interface that enables functions provided by an OS or middleware to be used by other application software.

gaiyou/pdf/1s1s.pdf
[2] NTT DOCOMO: DOCOMO Open Innovation Cloud (In Japanese). https://www.nttdocomo.co.jp/biz/service/doic/
[3] NVIDIA: "Jetson AGX Xavier." https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-agx-xavier/
[4] NVIDIA: "Jetson AGX Xavier NX." https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/
[5] Telit: "FN980." https://y1cj3stn5fbwhv73k0ipk1eg-wpengine.netdna-ssl.com/wp-content/uploads/2021/02/Telit_FN980-FN980m_Datasheet.pdf
[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi: "You Only Look Once: Unified, Real-Time Object Detection," Cornell University, May 2016. https://arxiv.org/abs/1506.02640

Technology Reports
MBH SR-MPLS TI-LFA

SR-based Routers in 5G MBH

Radio Access Network Development Department
Yoshio Igaue Takumi Harada
Yuki Matsuda Hiroaki Ogawa

In the 5G era, mobile networks have new requirements in addition to the conventional ones. It is becoming increasingly difficult for the MBH to meet the requirements for more efficient network construction and operation, higher reliability, large-capacity networks to meet the increased traffic in the 5G era, and for networks that can flexibly support a variety of services (low latency, guaranteed bandwidth, etc.) using only the connection method conventionally adopted for the MBH. Therefore, the MBH urgently needs further upgrades. As one solution, NTT DOCOMO has been considering a network design with a new connection method using SR technology. This article provides an overview of the newly commercially introduced SR-based router and its elemental SR technology.

1. Introduction

On a mobile network, the section connecting base stations and the core node*1 is called the "Mobile BackHaul" (MBH). The MBH needs to efficiently connect the base stations installed over wide areas to expand the areas available to users to the core node. Although the conventional MBH has had network design for excellent efficiency in traffic accommodation and reliability in case of device failure, with the coming of the 5th Generation mobile communications systems (5G) era, the MBH will require new design to cope with increased traffic and achieve 5G requirements for high speed and large

*1 Core node: A node such as an exchange or subscriber information management unit.

capacity (eMBB: enhanced Mobile BroadBand), low latency (URLLC: Ultra Reliable and Low Latency Communications), and massive terminal connectivity (mMTC: massive Machine Type Communications). To efficiently accommodate base stations in the MBH for these new requirements, it is assumed that issues below will be difficult to address with the equipment for the Ethernet method [1] that NTT DOCOMO has been operating so far.

(1) Providing a high-capacity network to meet the increased traffic in the 5G era (existing equipment has a maximum of 10GbE (Gigabit Ethernet))
(2) Improving efficiency of network construction/operation and further enhancing network reliability
(3) Realizing routing to flexibly respond to various Service Level Agreements (SLAs)*2 (low latency/bandwidth guarantees, etc.)

To solve issue (1), it is necessary to expand the interfaces of the equipment itself. Therefore, we considered expanding the physical interfaces of the equipment from the current maximum of 10G to over 100G. To deal with issue (2), we considered the support for the Segment Routing (SR) method. To solve both issues (1) and (2), we introduced a new MBH system using SR-based routers. To solve issue (3), we are also considering how to achieve routing that satisfies SLAs suitable for various services across the entire mobile network and to improve the efficiency of the network and operations by collaborating with the SR controller*3.

This article introduces an overview of SR-based router commercially deployed since 2020, the network configuration of an Access Router Network (hereinafter referred to as an "ARN"), which is a new MBH using SR-based routers, a functional overview and advantages of SR, which is the elemental technology that comprises an ARN, and the fault detection functions of the equipment. Finally, the article describes the future functional enhancement of the MBH.

2. Devices Comprising an ARN and Network Design

2.1 ARN Overview

An ARN consists of two types of Network Elements (NEs)*4, one called an "Area aGgregation Router" (AGR), the other called an "Area aCcess Router" (ACR), and an NE Operation System (NE-OpS)*5 which is responsible for monitoring, controlling, and configuring the NEs. An overview of an ARN configuration is shown in **Figure 1**.

We can install multiple line cards*6 with 100GbE and 10GbE or 1GbE physical interfaces in AGRs/ACRs, and construct facilities to guarantee transmission capacity according to the required bandwidth for the network using a Link Aggregation Group (LAG)*7 (**Table 1**).

ACRs are located near base stations and are responsible for aggregating multiple base station areas and their communications, and connecting them to AGRs. When ACRs connect to AGRs, the network topology is not only the star*8 type, but also enables ACRs to connect directly to neighboring ACRs without going through AGRs, which avoids the need to construct long-distance transmission lines through buildings. This reduces the cost of transmission lines and enables the network topology to be flexibly and easily expanded when

*2 SLA: A guarantee of the quality of a provided service.
*3 Controller: A device responsible for controlling NEs (see * 4).
*4 NE: A generic term for equipment that makes up a network.
*5 NE-OpS: A general term for a system that monitors and controls NEs.
*6 Line card: A module installed in a chassis-type router device, on which physical interfaces are installed.
*7 LAG: A function that treats multiple physical circuits as a single virtual circuit.
*8 Star topology: A type of network topology. A network structure in which multiple communication devices are connected to a central communication node.

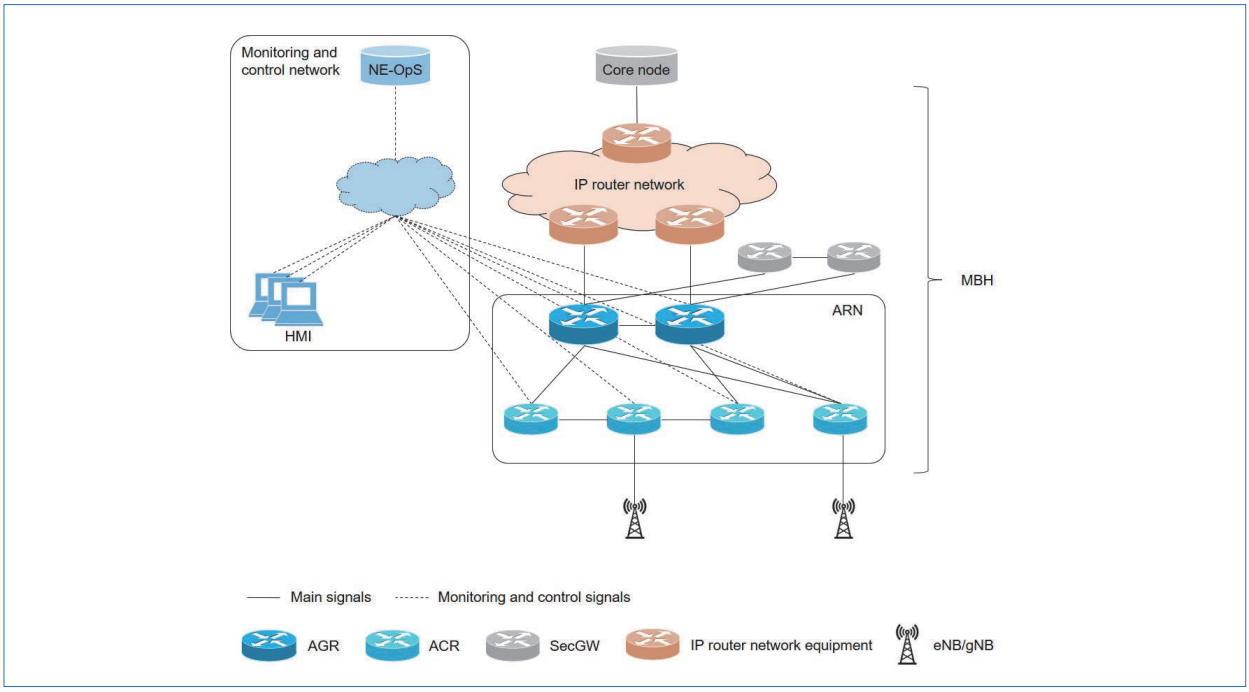


Figure 1 ARN configuration

Table 1 ARN equipment

	AGR	ACR
Chassis		
Line card (100G module)		

incorporating new ACRs. Also, when connecting to AGRs, by making physical lines redundant as necessary, we can construct a reliable network.

AGRs are routers that aggregate communications from the base stations constructed in each area and connect them to the IP router network [2]. Since high reliability is required, these routers are made redundant as network topology, which allows communications to continue without affecting services even if a router fails. As well as

aggregating communications, they also connect to Security GateWay (SecGW)*9 for encrypting communications in the backhaul section and the monitoring and control network.

A NE-OpS is connected to NEs via a monitoring and control network, which allows remote monitoring, control and configuration of all NEs.

2.2 ARN Characteristics

1) Flexibly Designable According to Various User and System Requirements

On an ARN, by applying Virtual Private Networks (VPNs)*10 to system units such as cellular systems including 5G, and wireless LAN systems, various services can be accommodated on the same physical network. This allows optimization of routing, Quality of Service (QoS)*11, security and latency for each user, and also ensures operational independence among users.

2) Route Aggregation

In a cellular system, a huge amount of routing information is distributed between an ARN and the IP router network, which is the upper layer network. Therefore, AGRs, which are the connection devices to the upper layer network, aggregate the routing information and broadcasts it to the upper layer (IP router network) to reduce the amount of routing information entered into the routing table maintained by AGRs and the IP router network. This reduces the load when resetting the routing table, learning the routing information again and entering it into the routing table with increased routing information on the entire network and thus enables stable network operations when a failure occurs.

3) Shaping*12

Since the base stations connected to an ARN cannot guarantee enough bandwidth (there are many cases using 1GbE connection when connecting a base station and an ARN), the shaping target downlink traffic coming from the IP router network is flagged at the AGRs, which are the entrances to the ARN, and shaping is applied to the flagged traffic at the ACRs, which are the exits from the ARN, to adjust the transmission bandwidth according to the transmission path. Even for narrow-band sections, this makes it possible to reduce congestion*13 and its resulting packet loss by reducing the traffic load in the ARN.

4) micro-Bidirectional Forwarding Detection (micro-BFD) and Border Gateway Protocol-Prefix Independent Convergence (BGP-PIC) Functions

In the event of network failure, in addition to redundant routers as described above, an ARN can be configured using technologies such as micro-BFD and BGP-PIC, described below, to quickly complete the process from failure detection to rerouting.

(a) micro-BFD

The AGR to AGR links, AGR to ACR links, and AGR to IP router network links are connected with multiple 100GbE interfaces for line redundancy, and are operated as one logical interface by applying a LAG. This function sends and receives keepalive packets at a high frequency on each physical interface that constitutes the LAG, which enables fast rerouting by detecting failures promptly.

(b) BGP-PIC

Since there are tens of thousands of routes

*9 SecGW: A communication device that terminates encrypted communication to/from a base station

*10 VPN: A virtual network logically structured for each service.

*11 QoS: A quality specification on a network for controlling the bandwidth used, the amount of latency and the discard rate, etc.

*12 Shaping: A type of bandwidth limitation method. A technology that buffers and transmits data when the data volume exceeds the threshold set in the device.

*13 Congestion: Impediments to communications services due to communications requests being concentrated in a short period of time and exceeding the processing capabilities of the service control server.

broadcasted in a cellular system, it takes seconds to find a backup route (lookup) in the event of a failure. BGP-PIC improves the calculation efficiency of the backup route lookup, and the time lag between failure detection and rerouting can be shortened.

2.3 Operation and Maintenance of an ARN

The NE-OpS manages, monitors, controls and configures the NEs. The system achieves efficient maintenance and operations through centralized management of ACRs and AGRs located in many locations. Communications between NE-OpS and NEs is via the ARN. As mentioned, the design minimizes the impact on operations even if a failure occurs in the ARN, and the network configuration has little impact on management and monitoring of the NEs (ACRs/AGRs), so that maintenance and operations can be continued even if a failure occurs.

3. SR Functions

This chapter provides an overview of the main SR functions operating on an ARN.

3.1 SR Overview

In general, router devices determine the next-hop router based on the destination IP address of the packets. In case of SR, the next-hop is determined with "segment". This architecture enables simpler and more flexible control than the conventional Multi-Protocol Label Switching (MPLS)[*14]-based forwarding. While there are several types of segment, an ARN mainly uses the Interior Gateway Protocol (IGP) – Prefix Segment and IGP-Adjacency Segment. An overview of the SR forwarding operation is shown in **Figure 2**.

1) Prefix-SID (IGP-Prefix Segment ID)

Prefix-SID is a unique value on the SR domain and is used for basic forwarding control. The Prefix-SID corresponding to the IGP-Prefix (e.g., loopback address[*15]) used for a single node is called the "Node (IGP-Node Segment) - SID". SR-based routers can use it to uniquely designate a node in a domain.

2) Adj (IGP-Adjacency Segment) - SID

This is a unique value within each SR router, is assigned to each link to neighboring IGP nodes, and is used in special forwarding control. An SR node forwards packets with its own Adj-SID through the corresponding link, without any IP shortest-path consideration.

3.2 SR Characteristics

An ARN uses the SR-MPLS, which applies SR to the MPLS data plane[*16] and has the following features.

1) Data Plane

In SR-MPLS the conventional MPLS data plane can be used as it is. The segment is also represented by encoding the SID in the Label field of the MPLS header.

2) Control Plane[*17]

The following three options are available to implement SR control plane.

- Distributed, in which the segments are allocated and signaled only by extended routing protocols (such as Open Shortest Path First (OSPF)[*18]). This method does not require label distribution-specific protocols (Label

*14 MPLS: A packet forwarding technology that uses labels instead of IP addresses.
*15 Loopback address: An IP address assigned to a node to indicate the node itself.
*16 Data plane: A routing table and its logic used by routers to forward traffic.
*17 Control plane: A routing protocol used by routers to control traffic forwarding.
*18 OSPF: A protocol that allows routers to select routes based on adjacent connection information.
*19 LDP: A kind of label exchange protocol. Label information exchange enables the establishment of a session between NEs.

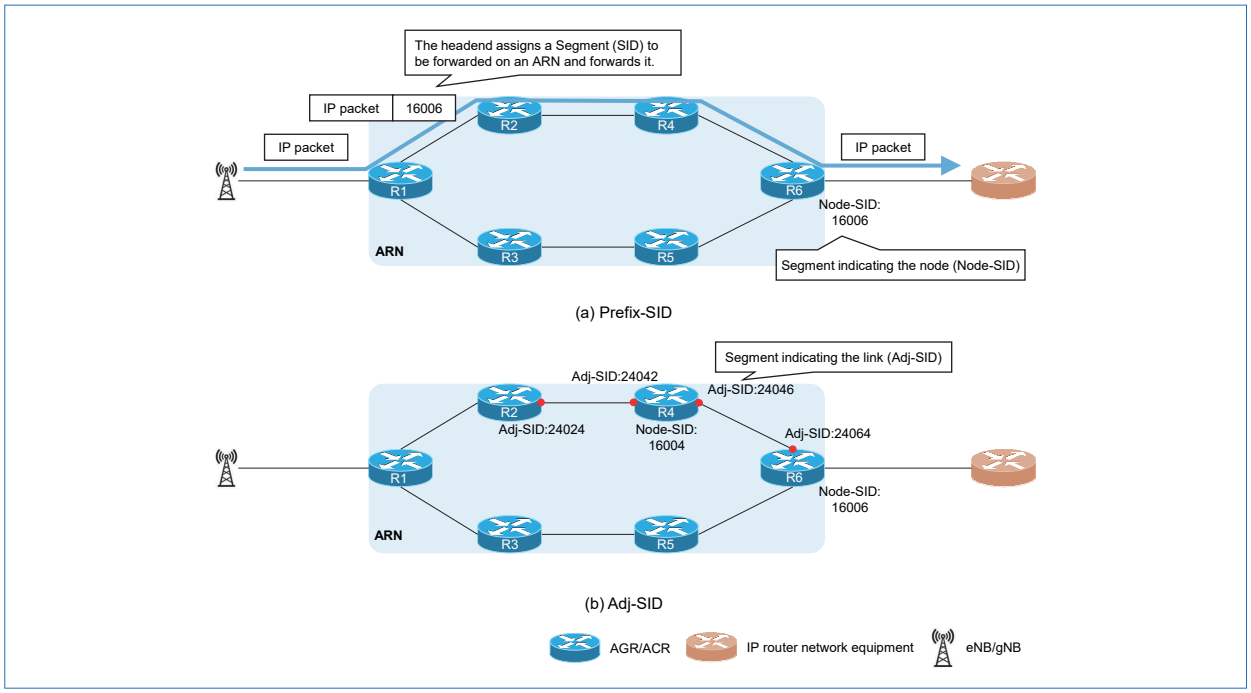


Figure 2 Overview of SR transfer operation

Distribution Protocol (LDP)*19 or Resource reSerVation Protocol (RSVP)*20), making it simpler than conventional MPLS (OSPF uses this method in an ARN).

- Centralized, in which the segments are allocated and instantiated by the SR controller.
- Hybrid, which is based on the distributed control plane with the SR controller

3) Source Routing

An ingress SR node can steer a packet through an SR-policy*21 instantiated as an order list of instructions. A transit SR node determines the next-hop following the SR-policy. In addition, the architecture is highly compatible with a Software Defined Network (SDN)*22 (enabling easy centralized control with a controller) and can provide optimized routing operations for each user, such as low latency and guaranteed bandwidth.

4) Fast Failover

Since an ARN deals with the C-Plane*23 of cellular communications and voice services, it requires a high level of availability. An ARN must be designed to minimize packet loss with quick rerouting in the event of any failure. As shown in **Figure 3**, in conventional MPLS, there is a concern that, depending on the topology, a packet rerouted once may end up heading toward a failure point again, resulting in a loop (Fig. 3(b)). However, in SR-MPLS, loops can be avoided by strictly specifying the forwarding backup route using the MPLS label stack (Fig. 3(c), (d)). This is called "Topology Independent Loop Free Alternate" (TI-LFA), and is a function that enables fast rerouting for failures on the IGP (OSPF) network. Conventional IP fast reroute such as LFA and remote LFA (rLFA) cannot install

*20 RSVP: A protocol to secure necessary QoS for services such as data requiring real-time communication by reserving in advance end-to-end bandwidth on an IP network.
*21 SR-policy: A policy that specifies a forwarding route on an SR network.
*22 SDN: Software control of the configuration and operation of devices comprising a network.
*23 C-Plane: A protocol for transferring control signals for establishing and disconnecting cellular communications.

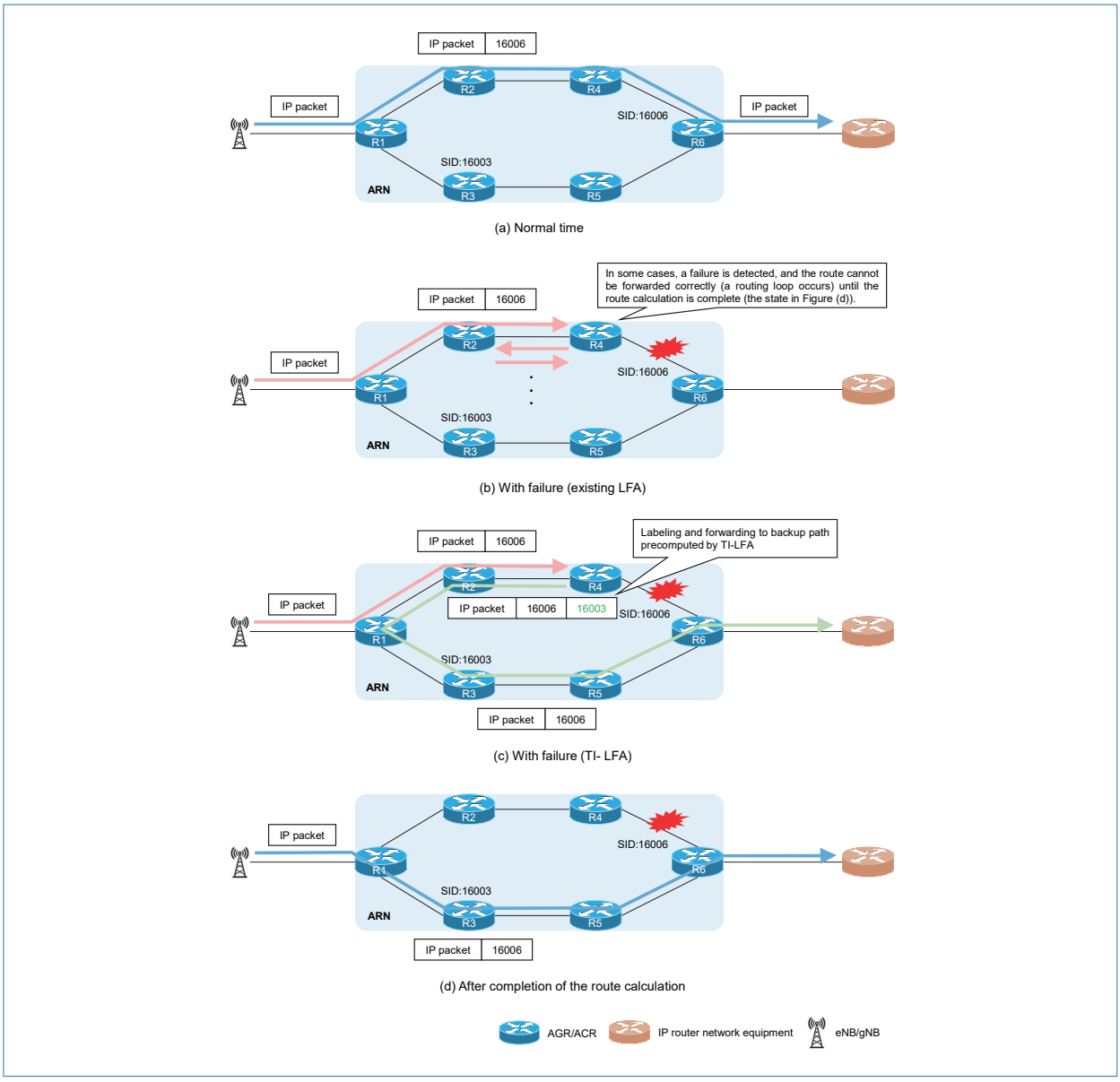


Figure 3 Summary of failure actions and TI-LFA actions on the ARN

a loop-free reroute depending on the topology.

5) Operation And Maintenance (OAM)*24 Functions

SR-based routers have several OAM functions.

First of all, as basic OAM functions, ping*25 and traceroute*26 can be used to check reachability to neighboring nodes within an IGP network.

Next, Virtual Routing and Forwarding (VRF)-ping/traceroute can be used to check the reachability and forwarding route to any destination on the VPN. The destination does not have to be on the ARN, and the function can also be used to

*24 OAM: Functions for maintenance and operational management on a network.

*25 Ping: A command to check the reachability of a node on an IP-network.

*26 Traceroute: A command to check the network route to a node on an IP-network.

check the reachability to base stations and core nodes.

In addition, MPLS-Ping can be used to check reachability of an IGP network, etc. on the MPLS data plane. This is used to check for special issues such as failure to exchange SID information.

One of the features of SR-based routers is Data Plane Monitoring (DPM), a function that detects silent failures*27. As shown in **Figure 4**, DPM enables periodic checking of the liveness of the network.

4. Evolution of the MBH in the 5G era

4.1 Overview of SR-TE

Applying SR technology to Traffic Engineering (TE), which can explicitly specify routes in an ARN, makes it possible to achieve even more flexible routing to meet each service requirement. Similarly, in the MBH, SR-TE can provide highly detailed network control to meet the various needs of users. On the mobile network, communications from user mobile phones to the core node are the target of SR-TE in our network.

With the launch of 5G services, a variety of requirements with lower latency, higher capacity, and more simultaneous connections to many devices than ever before must be accommodated. We require even more flexible network design that considers the characteristics of each network providing these services.

An overview of SR-TE is shown in **Figure 5**. In SR-TE, the route of packet forwarding is determined by source routing as described in Chapter 3 so

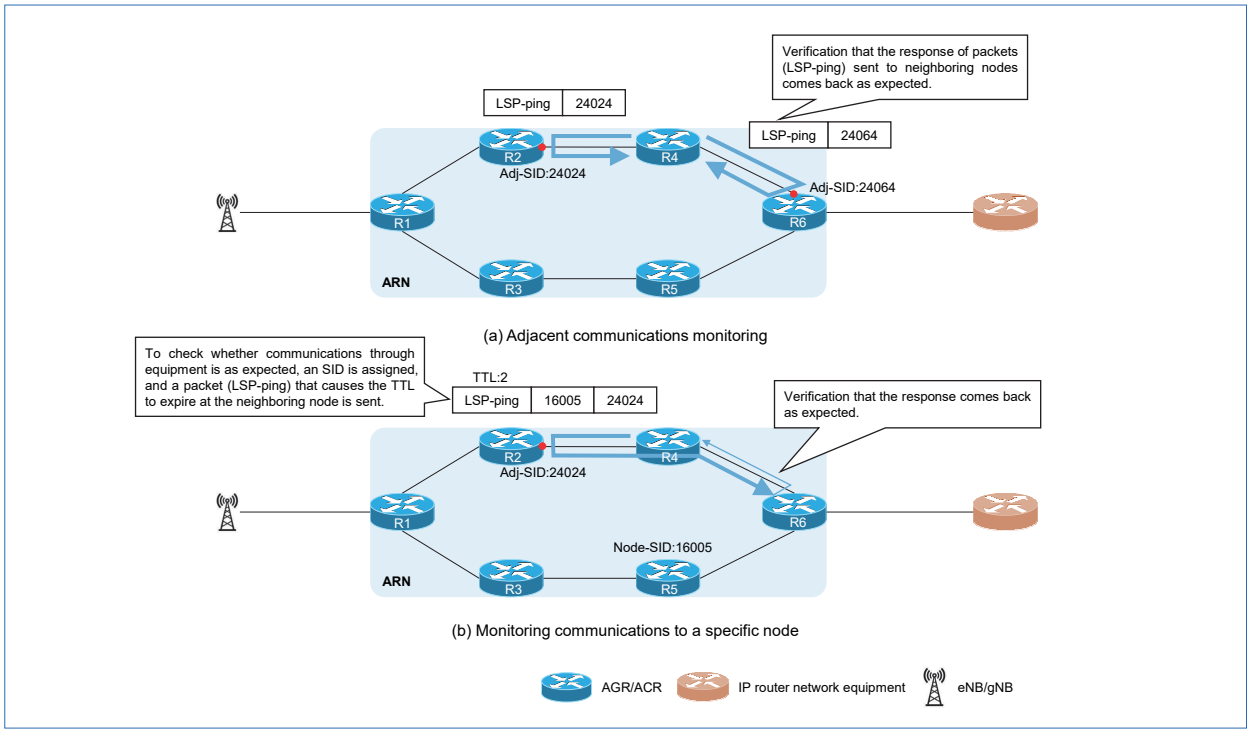


Figure 4 Overview of DPM operations

*27 Silent failure: A failure that cannot be detected by NE-OpS.

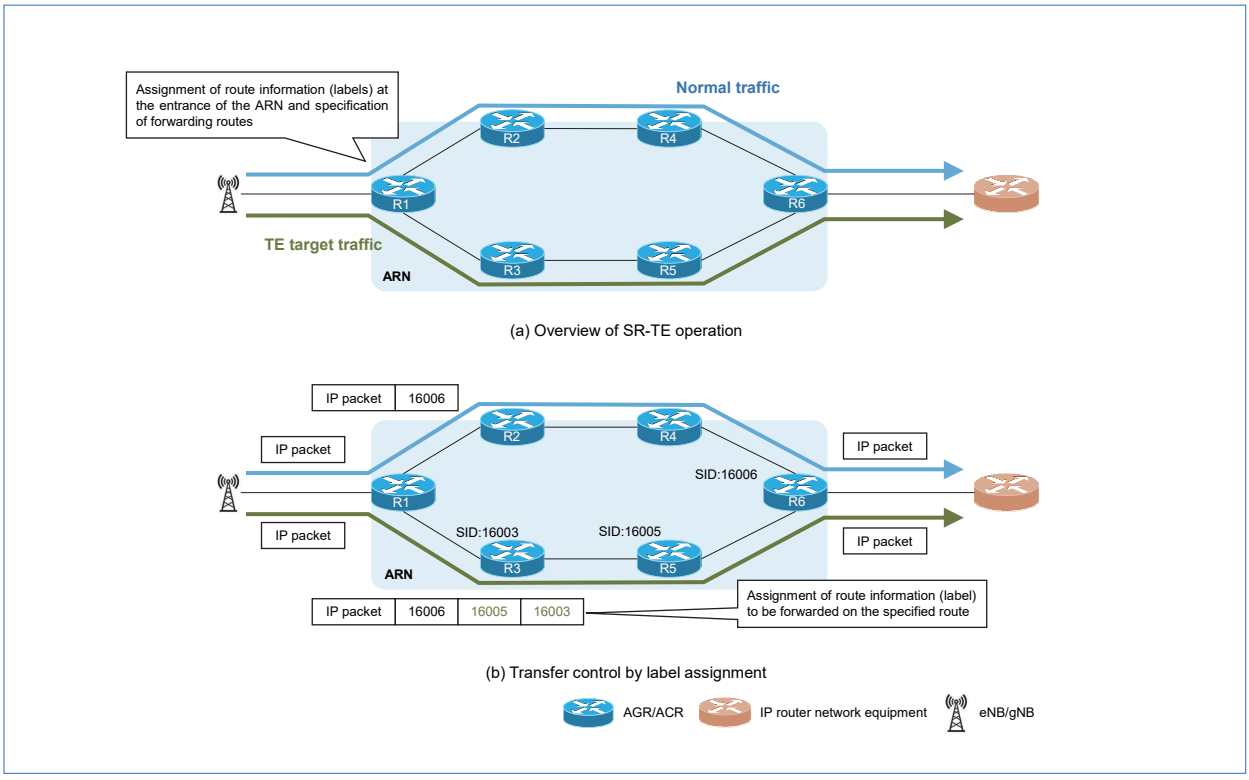


Figure 5 Overview of SR-TE control

that the route through the MBH can be explicitly specified at the entrance to an ARN. When packets flow into the MBH, route of packet forwarding is specified using the identifiers assigned to each service, making it possible to operate routes as logically separate networks. For example, IP addresses, port numbers, and Differentiated Services Code Point (DSCP) values*28 can be used as identifiers to sort services, and control methods can be tailored to the nodes to be connected.

For example, when providing low latency services, it is possible to calculate the forwarding path using the amount of delay between nodes and set the shortest path, which is the result of the calculation, to each piece of network equipment. It is possible to construct the MBH more flexibly than in the past by using routes determined by route calculation using path cost*29 such as OSPF and routes determined by route calculation using the amount of delay.

4.2 Towards the Realization of Network Slicing in E2E

Designing the network through the SR controller will make it possible to optimize routes not only for the MBH but also for other networks connected to the MBH, which holds promise for the realization of End-to-End (E2E) flexible routing. Slicing in E2E is expected to provide users with networks that are even more in line with the requirements of each service unit.

The network topology through the SR controller

*28 DSCP value: A value indicating priority level when controlling the priority of packets.

*29 Path cost: The accumulated distance (weighting) on the path to the destination.

is shown in **Figure 6**. Since the SR-based router alone cannot collect the network configuration information other than the ARN, it is necessary to obtain the network configuration via the other controller for the destination network. Path computation is not performed only by the routers on the ARN section and the SR controller for the ARN, but also in cooperation with the other network controller, sharing other network configuration information necessary for routing and performing path computation.

Coordinating between each controller and configuring each router under each controller in this way makes it possible to apply a consistent routing policy in E2E and configure logical networks to meet various service requirements. In this case, the SR-based router will realize routing in cooperation with the controller, making it possible to provide flexible network design for a variety of services and requirements.

5. Conclusion

This article has described an overview of the SR-based router to support the 5G era, its network configurations, and MBH evolution that will be realized in the future.

As discussed, the introduction of the SR-based router enables a significant increase in network capacity and efficient network bandwidth utilization, and also increases network reliability in the event of failures due to fast failover and improved failure detection.

Going forward, we will continue to consider the enhancement of functions to flexibly support a wide variety of 5G services through linkage with

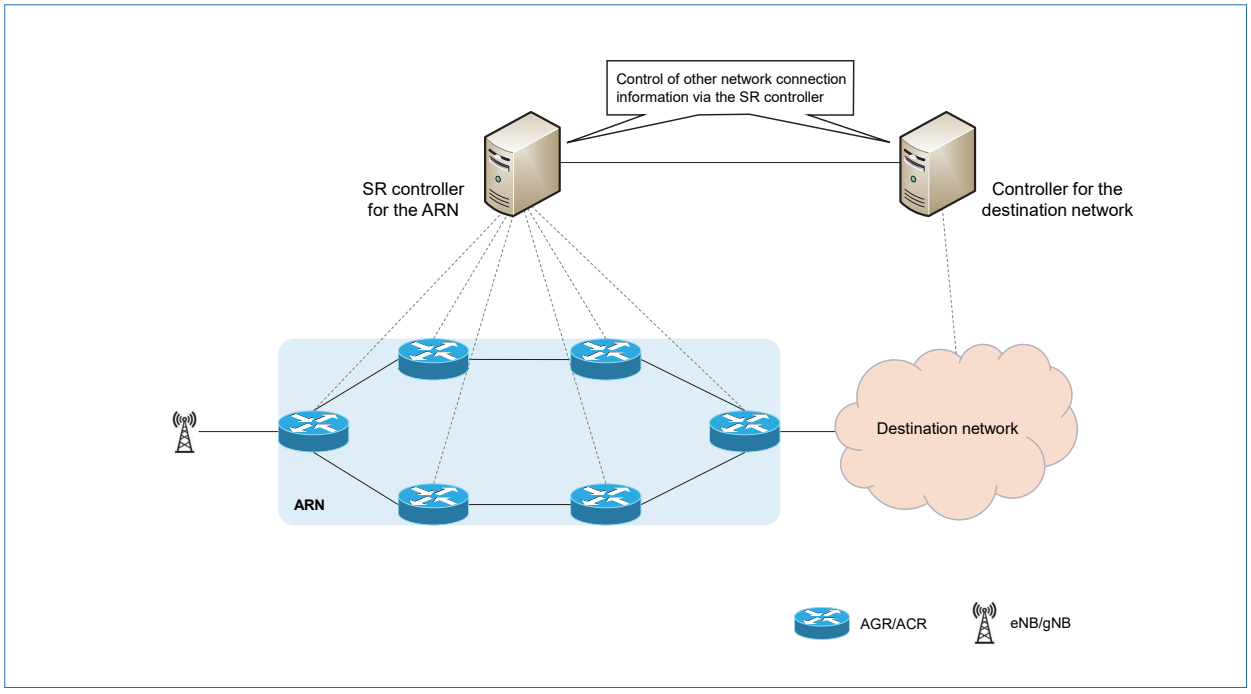


Figure 6 Network topology through the SR controller

the SR controller.

REFERENCES

[1] K. Morita, et al.: "Ethernet Transmission Equipment ERP-SW for All-IP Transmission Paths," NTT DOCOMO Technical Journal, Vol.12, No.3, pp.56–60, Dec. 2010.

[2] K. Osaki, et al.: "Configuration Technology for "IP Router Network" – Overview of L2/L3 Integrated VPN Backbone –," NTT DOCOMO Technical Journal, Vol.12, No.1, pp.57–63, Apr. 2004 (In Japanese).

Technology Reports Network Virtualization Orchestration Upgrade without Service Interruption

Migration to ETSI NFV Stage 3 Specification-compliant Multivendor MANO Configuration on Network Virtualization Platform

Core Network Development Department Hironao Tamura Yuya Kuno
DOCOMO Technology, Inc. Yoshiyuki Suzuki

NTT DOCOMO commercially introduced a virtualization platform for its core network at the end of FY2015, and has continued to operate it stably since then. This virtualization platform supports multivendor configuration and also is able to accommodate the operation of many virtualized core network functions running on it. To address the EoL of NFVO and VNFM, which are components of the virtualization platform, we have developed a multivendor configuration framework over the virtualization platform that incorporates the latest ETSI NFV Stage 3 specifications. Also, we succeeded with the migration from the legacy virtualization platform to the new one without service interruptions or negative impacts on communications services. This article describes these activities.

1. Introduction

At the end of FY2015, NTT DOCOMO became the first company in the world to configure a virtualization platform on which a core network*1 is running in an operational environment consisting of products from multiple vendors. This reduces communications network CAPital EXpenditure (CAPEX)*2/OPerating EXpense (OPEX)*3 [1] [2]. Since then, NTT DOCOMO has continued to virtualize more core network functions and operate the virtualization platform stably. By the end of

*1 Core network: A network comprised of switching equipment, subscriber information management equipment, etc. A mobile terminal communicates with the core network via a radio access network.
*2 CAPEX: The amount of money expended for hardware and hardware installation.
*3 OPEX: The amount of money expended for maintaining and operating facilities.

FY2020, more than 50% of these core network functions had been virtualized, with more than 400 running on the virtualization platform. The lifecycles of these virtualized core network functions (Virtual Network Functions (VNF)*4) are managed by Management ANd Orchestration (MANO)*5. Lifecycle management includes automatic healing to recreate Virtual Machines (VMs)*6 on normally operational hardware if a hardware or VM failure occurs, scaling to optimize processing power by increasing or decreasing processing resources according to the load on the hardware and VMs, and VNF upgrading to update the communications software and VM-related software when adding new services or solving issues. To provide stable communications services, this virtualization platform runs as a mission-critical*7 system (**Figure 1**).

As mentioned above, NTT DOCOMO's virtualization platform is composed of products from multiple vendors. These products differ in their End of Life (EoL)*8 timing, the version of standard specifications they support, their Application Programming Interfaces (APIs)*9, input/output parameters*10, and attributes*11. Each vendor decides these items based on standardization work and future trends, releases plans of them and specifications based on operator requirements. Therefore, when upgrading a system that is made by combining products from multiple vendors, the complexity of integration*12, including the design of InterFace (IF) specifications between products, is an issue. Addressing this, NTT DOCOMO actively participates in standardization at the European Telecommunications Standards Institute (ETSI)*13 Network Functions Virtualization (NFV)*14 [3] and has contributed to the development of Stage 3 specifications*15 [4]. We continue to work on standardization in consideration of integration, reliability, and

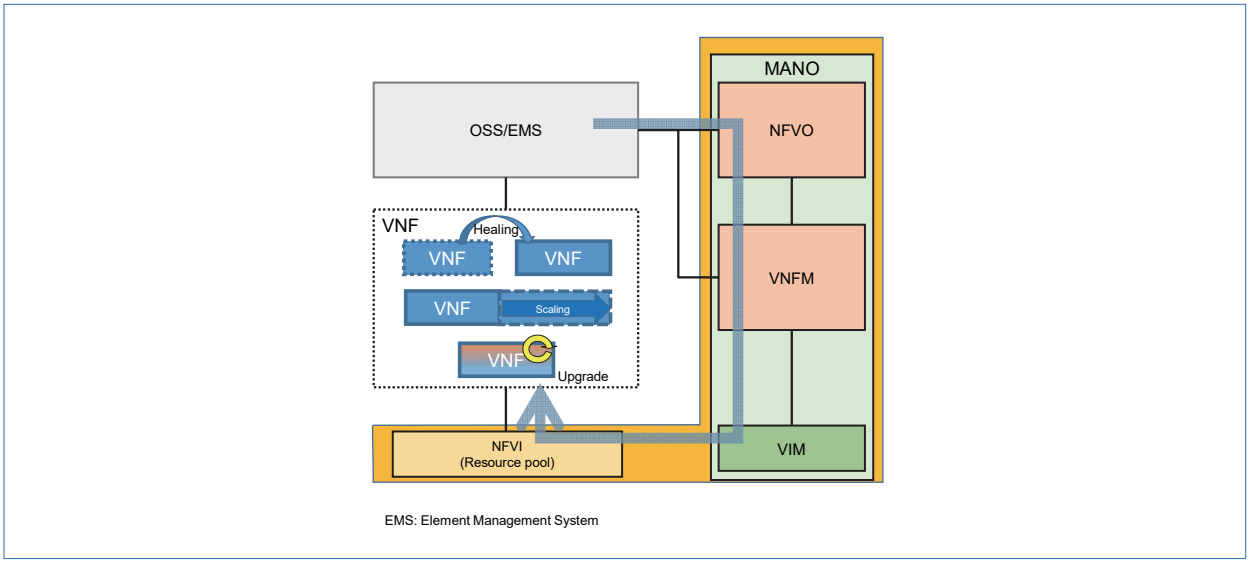


Figure 1 Healing, scaling, and VNF upgrading on the virtualization platform

*4 VNF: Virtualized communications functions (communications systems).
*5 MANO: A generic term that summarizes the functional blocks of NFVO (see *16), VNFM (see *17), and VIM (see *18).
*6 VM: A computer (virtual machine) created in a virtual manner by software.
*7 Mission-critical: Refers to elements essential to the execution of businesses or services for which termination or interruption is not allowed.
*8 EoL: Refers to cessation of product manufacture and sales, cessation of software product support services, or cessation of provision of modification/update programs for bug fixing and functional improvements.
*9 API: A specification for exchanging information for connecting software components and products.
*10 Parameter: A generic term for a set of data exchanged between systems in ETSI NFV. When using an API, the data passed by the system that calls the API is called the input parameter, and, as execution results, the data returned by the system that opens the API and executes the API is called the output parameter.

operational efficiency among multivendor products.

In this article, we describe our activities to coordinate IF specifications to comply with ETSI NFV Stage 3 to handle product EoL. The article also describes a method of upgrading systems without stopping communications services, a fundamental requirement of telecommunications providers. Systems to be upgraded are NFV Orchestrator (NFVO)*16 and VNF Manager (VNFM)*17, which are key components of MANO. However, since Virtual Infrastructure Manager (VIM)*18, which manages virtual resources in the cloud*19 environment, is not upgraded, we have to upgrade these systems while VIM is in service.

2. Virtualization Platforms Landscape

2.1 Trends in Virtualization Platforms

NTT DOCOMO actively participates in standardization at ETSI NFV Industry Specification Group (ISG) and has contributed to the growth of NFV technologies and NFV standards adoption. As NFV becomes more widespread as a global standard, virtualization platforms are being deployed by many telecommunications operators. However, as well as single-vendor deployment, the type of deployment differs from operator to operator, with additional challenges including multivendor configuration aspects such as that of NTT DOCOMO. Furthermore, from the deployment perspective, since VNF architectures running on virtual platforms may differ, use cases and development methods for VNF LifeCycle Management (LCM)*20 can also vary. This means the way operators use ETSI NFV specifications and the scope of IFs virtualization platforms must support also differ.

2.2 Utilization of Global Products in a Multivendor Environment

In recent years, the application of cloud technologies has significantly changed the way operators develop their businesses. As cloud technology evolves more quickly and product lifecycles get faster, it is becoming less efficient to develop everything in-house. Meanwhile, it has been difficult to use systems other than those developed in-house for operation and maintenance, because each country has different legal frameworks and different geographical conditions, and operators must comply with different organizational structures and operational flows.

Therefore, as shown in **Figure 2**, it is extremely important to move from the legacy development processes based on an operator's own business analysis (an example of waterfall development) to new style that involves 1) the promotion of operator collaboration based on multifaceted business analysis considering the different operational workflows and needs of each operator, 2) the promotion and description of common requirements and development of common specifications through external standardization activities, and 3) the acceleration of development of standardized products that satisfy the needs of operators.

These activities will make products with uniform specifications globally available to many operators, thereby maintaining interoperability in a multivendor environment while lowering costs.

*11 Attribute: A generic term for resource information held by each functional block in ETSI NFV. Since the ETSI NFV specification is based on REST (see *36), the resources held by each functional block are called attributes, and are distinguished from input/output parameters in APIs.
*12 Integration: The incorporation of equipment or systems on a network operated by an operator.
*13 ETSI: The standardization organization concerned with telecommunications technology in Europe.
*14 NFV: Achieving a telecommunications carrier network on general-purpose hardware through virtualization technology.
*15 Stage 3 specifications: A general term for specifications of signaling protocols among the standardization methods in telecommunications specified in International Telecommunication Union-Telecommunication sector (ITU-T) I.130.
*16 NFVO: The system that performs comprehensive management of virtual resources that span multiple VIMs (see *18).
*17 VNFM: The system that controls VNF such as launching and terminating VNF LCM.

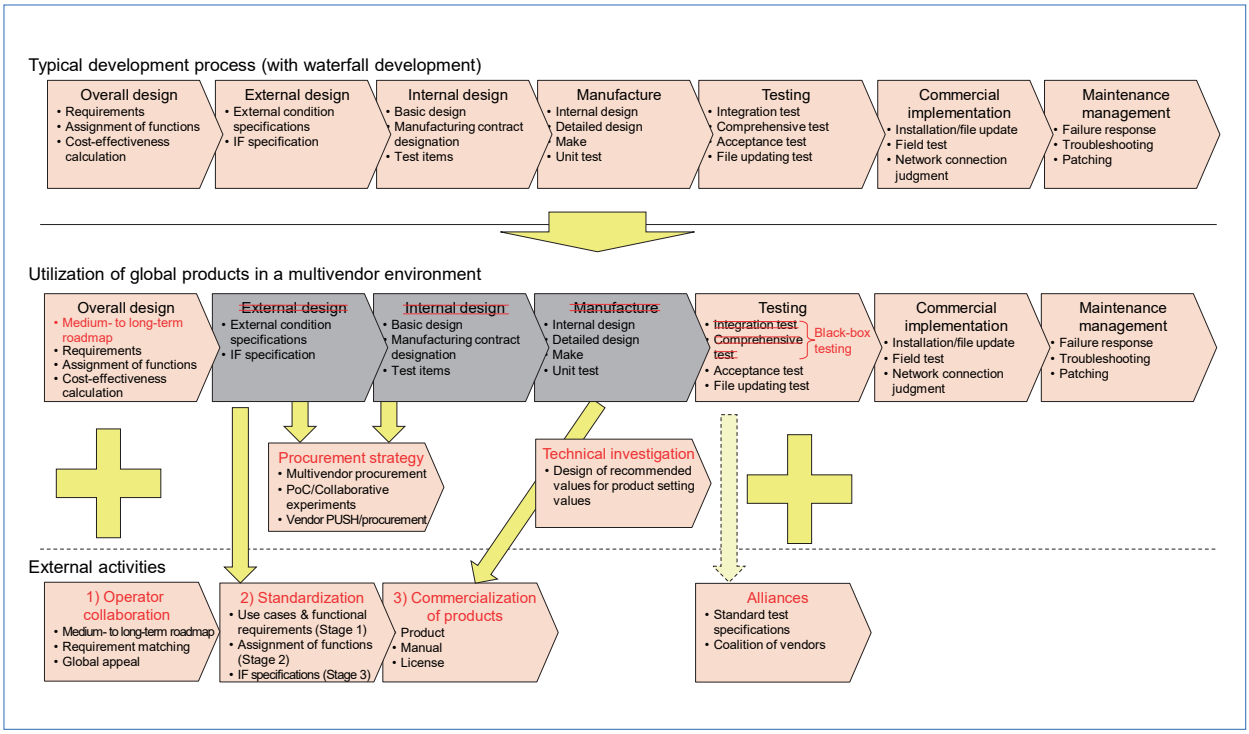


Figure 2 Utilization of global products in a multivendor environment

1) Operator Collaboration

Requirements need to be common among operators to ensure wide adoption of specifications. To achieve this, it is important for each operator to create its own medium- to long-term roadmap, and jointly coordinate requirements specifications with other operators (to build global appeal).

2) Standardization

Based on the common requirements specified by operators, use cases will be created jointly during the development of standards. By clarifying use cases, functional requirements will become clearer (Stage 1). Next, different functionality will be split into functional blocks and the information flows among them (Stage 2) can be specified. Finally, IFs (Stage 3) will be specified. Creating standard specifications among operators and vendors will make global products interoperable in a multivendor environment.

3) Promotion of Commercialization in Compliance with Standard Specifications

When operators procure products that comply with standard specifications from vendors, the products will become commercialized (product manufacture, manual creation and licensing).

Thus, there is no significant difference between development of virtualization platforms using global products and legacy development in terms of processes necessary for design, such as splitting

*18 VIM: The system that manages the physical resources on a virtualization platform consisting of physical computers, physical storage, and physical networks.

*19 Cloud: A format and mechanism for providing services over the network. Server resources can be allocated according to demand making for high scalability.

*20 VNF LCM: VNF lifecycle processing.

functions and establishing IF specifications, and product manufacturing and testing to verify that the product meets requirements. However, additional external activities are needed to develop common requirements and specifications.

3. NTT DOCOMO's Approach to MANO Upgrades

As mentioned above, since virtualization platforms require cloud technology, NTT DOCOMO has been working on standardization at ETSI NFV and has been promoting commercialization of products compliant with standard specifications while coordinating IF specifications for multivendor MANO configuration, and has been developing a framework for upgrading without interrupting services.

3.1 Identifying De Facto Versions and Promoting Commercialization

Communications equipment must continue to provide services for a period of 10 years, but as mentioned above, the adoption of cloud technology tends to shorten product lifecycles and frequently involves changes to the functions themselves. As the goal is to put communications equipment in the cloud, ETSI NFV is moving closer to a cloud computing culture, and its standard specifications will have a minor version upgrade every six months and a major version upgrade every two years with function addition and changes to the roles of functions occurring each time.

Therefore, when connecting multiple products, it is necessary to determine which version of which specifications is supported and the support period for that version, based on the assumption that operator use cases and requirements have been standardized through the aforementioned 1) Operator collaboration and 2) Standardization efforts. In addition, since supported functions and security-related specifications may change with each version, an operator must be able to ask vendors for specific versions that meet its requirements and cover its use cases.

While identifying de facto versions based on these considerations, it will be important to consider the aforementioned 3) Promotion of commercialization in compliance with standard specifications, provide education about de facto versions in an open forum, demand these de facto versions from product vendors and align functional requirements among operators. Therefore, it is important to raise the incentive for each product vendor to commercialize products in terms of what versions of what products are needed and by when.

3.2 Study of Methods to Start Upgrade Developments

IFs that comply with ETSI NFV Stage 3 are shown in **Figure 3**. Vendor products are basically commercialized based on market needs. Unlike communications equipment such as equipment that supports roaming[*21] which cannot provide communications services unless standard specifications are fully implemented, standard specification are not necessarily implemented with operation and maintenance equipment such as Operating Support System (OSS)[*22]/Element Manager (EM)[*23] and MANO.

*21 Roaming: A mechanism that enables mobile subscribers to receive services when visiting the service areas of partner telecommunications carriers, when traveling outside the geographical coverage area of their home networks.

*22 OSS: An operation support system for network operations for discovering, controlling, and dealing with faults and congestion in a mobile communications network. For a network operator, this means full or partial network or system fault management, configuration management, billing management, performance management and security management for usage of provided services.

*23 EM: A system that manages and monitors Faults, Configuration, Accounting, Performance and Security (FCAPS) for individual pieces of communications equipment.

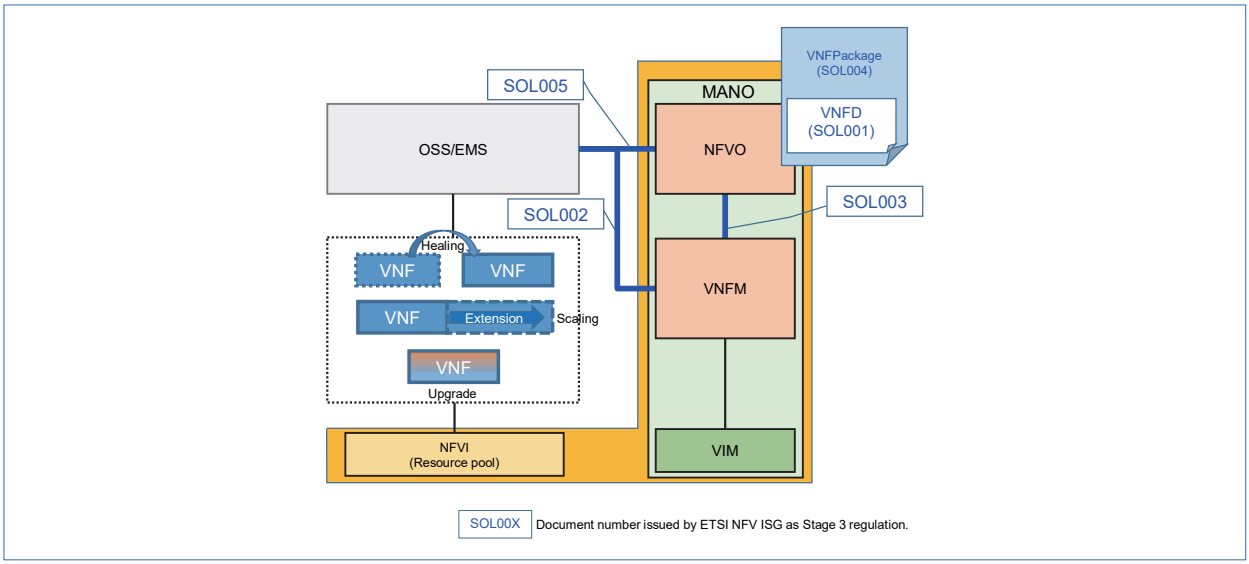


Figure 3 ETSI NFV Stage 3 IF

In the case of operation and maintenance equipment, the functions required are completely different depending on the operator's workflows, so vendors implement only the essential common functions and IF specifications requested by the majority of operators in their products. As a result, when connecting different MANO products, even if the IF specification is defined as a standard specification, each product might support different APIs with different attributes and values*24. For this reason, a feasibility study to connect MANO products is required.

To ensure compliance with standard specifications, it is necessary to complete the use cases of a company as soon as possible, as described in 1) to 3) below, and to clarify and request from the vendor the minimum necessary functions compliant with standard specifications for the deployment of new versions of equipment. In addition, if an operator requires implementation of non-standardized functions or extensions*25 in products to realize a use case, it should not prevent stable upgrades, and the operator should carefully consider selecting interoperable versions to avoid problem solving difficulties during the testing process or commercial operations. Therefore, it is important to have strong policies to prevent non-compliance with standard specifications to maintain interoperability between products. Towards the development of version upgrades, the following is a list of considerations to achieve these goals.

1) Create a Product-to-product Sequence*26 and Determine the Assignment of Functions in New Versions

Sequences using standard specifications must be considered based on a company's own use cases. The major types of sequences are (1) deployment*27 (onboarding*28 and instantiation), (2) update

*24 Value: The specific value passed by an attribute. In REST (see *36), this is passed by a key value pair consisting of an attribute and a value.

*25 Extension: Parameters for APIs and VNFDs (see *35) that have been extended to allow products to use their own functions.

*26 Sequence: The order of processing agreed upon between systems, functional blocks, etc.

*27 Deployment: Installing applications by placing them in their execution environments.

*28 Onboarding: Registering a VNF Package in the MANO system.

(healing, scaling and file update), and (3) removal (termination*29). To achieve these goals, it is necessary to determine products that will take on the role of VNF Instance*30, Virtual Resource*31, and VNF Package*32 resource owner*33, and to create the sequence.

2) Determination of Error-handling*34 Methods and Manual Recovery Procedures Based on Assignment of Functions

ETSI GS NFV-SOL 003 specifications describe policies and specifications for error-handling, although the supported functions and implementation-dependent error cases need to be well investigated. In particular, it is necessary to confirm detection and recovery methods to address communication failures - whether to recover using the retry method or the rollback method for each use case - and the error-code and recovery method when VNF LCM fails.

3) Determination of Parameter Values to Use

In ETSI NFV standardization, many parameters and attributes have been defined to support different use cases, and many options regarding information flows are available. Therefore, based on the sequence in 1) and the error-handling method adopted in 2), it is necessary to determine how main parameters are exchanged among multiple systems, what should be described in the VNF Descriptor (VNFD)*35, and which parameters and attributes are necessary to realize a use case. In particular, it is necessary to determine the network connectivity information between VNFs, the placement policy for VMs to form a redundant configuration to maintain reliability, the identifiers of VIMs and VNFs used in virtual resource management, and naming rules so that each piece of equipment can uniquely determine these parameters.

Through these studies, it is important to document the sequence, error-handling, each parameter value, etc., and request them from each vendor as mandatory features of each product. To implement future upgrades smoothly, NTT DOCOMO has contributed to the standardization of sequences (procedures), parameter exchange (key information exchange) between pieces of equipment, competitive conditions (dependent and non-dependent side procedures) in REpresentational State Transfer (REST)*36 environments and main error-cases as ETSI GS NFV-SOL 016.

3.3 Upgrade without Service Interruption

More than 400 core network VNFs run on NTT DOCOMO's virtualization platform. Any negative impact on these functions will directly lead to degradation of user services, such as connection difficulties or suspension of user communications. Therefore, when upgrading NFVOs and VNFMs, the upgrade must not affect the communications services provided by operational VNFs .

The following are requirements for upgrading NFVO and VNFM without impacting communications services.

(1) No need to recreate the VNF

One possible upgrading method is to create a new VNF (instantiation) on a new system built separately from the old system, and switch to the newly deployed VNF. However,

*29 Termination: Stopping a VNF and removing it from a virtualization platform.
*30 VNF Instance: A VNF entity created (instantiated) on a virtualization platform by a MANO system.
*31 Virtual Resource: Virtualized hardware resources (CPU, memory, hard disk, etc.).
*32 VNF Package: A package containing VNFD (see *35) that defines VNF functions and operations, ancillary information and VNF images.
*33 Resource owner: The system from which VNF instance, virtual resources and package resources are managed.
*34 Error-handling: Responses when errors occur.
*35 VNFD: The definition file for VNF to be created on the virtualization platform.
*36 REST: An API that sends GET, POST, PUT, and DELETE requests to each resource (URL) and receives the response in XML, json, or other format (the format of the response is not specified).

with some core network functions, communications services may be impacted when switching VNFs, and recreating and switching more than 400 VNFs requires a huge period of time and cost. Therefore, it is preferred that VNFs must not have to be recreated.

(2) The lifecycle management for the VNF that is running should be executable (stopping for a certain period of time is acceptable).

On a virtualization platform, healing is a type of lifecycle management which improves reliability when a hardware failure occurs. Specifically, when failures occur, VMs in which VNFs are deployed are recreated on other normally operating hardware to maintain redundancy. If healing is not possible, VNF redundancy cannot be maintained, and the risk that the next hardware failure will affect the communications service provided by VNFs is high. Therefore, it is necessary to be able to continuously perform healing.

(3) The upgrade process must be reversible.

It is necessary to be able to revert an upgrade and restore normal operations when abnormalities occur during the upgrade.

We upgraded the NFVO and VNFM by meeting these requirements and not impacting communications services. The mechanism is described below.

1) Mechanism to Eliminate the Need for VNF Recreation

To provide VNF LCM functions, NFVOs and VNFMs maintain management data about the virtual resources of the VNFs that they run (hereinafter referred to as "VNF data"). To eliminate the need to recreate VNFs when upgrading, it is necessary to move the VNF data managed in the old system to the new system, so that VNFs can continue to run on the new system.

In general, there are two possible ways to upgrade a system that has inherited data: update an old system to a new system on the same server, as in an OS update, or build a new system on a different server from the old system and switch to it after data migration. Since the hardware requirements of the target NFVO and VNFM were changed in this upgrade, a new system was built on a separate server and switched after data migration. The NFVO manages multiple VNFMs, including VNFMs that are not upgraded. The upgrade was performed in two stages, NFVO and VNFM, as shown in **Figure 4**.

As for data migration, both NFVO and VNFM will need to be converted to the new data format when VNFM is upgraded in the second stage. Since the upgraded NFVO and VNFM will be compliant with the ETSI NFV Stage 3 specifications, they will involve some changes in the assignment of functions from the current system. The structure of the VNF data to be managed and the parameters of the inter-system IF will be significantly different. Therefore, we found that migrating data from the current system to the new system required complex and extensive data conversion and harmonization across NFVO and VNFM (**Figure 5**). The development of these processes involves a risk

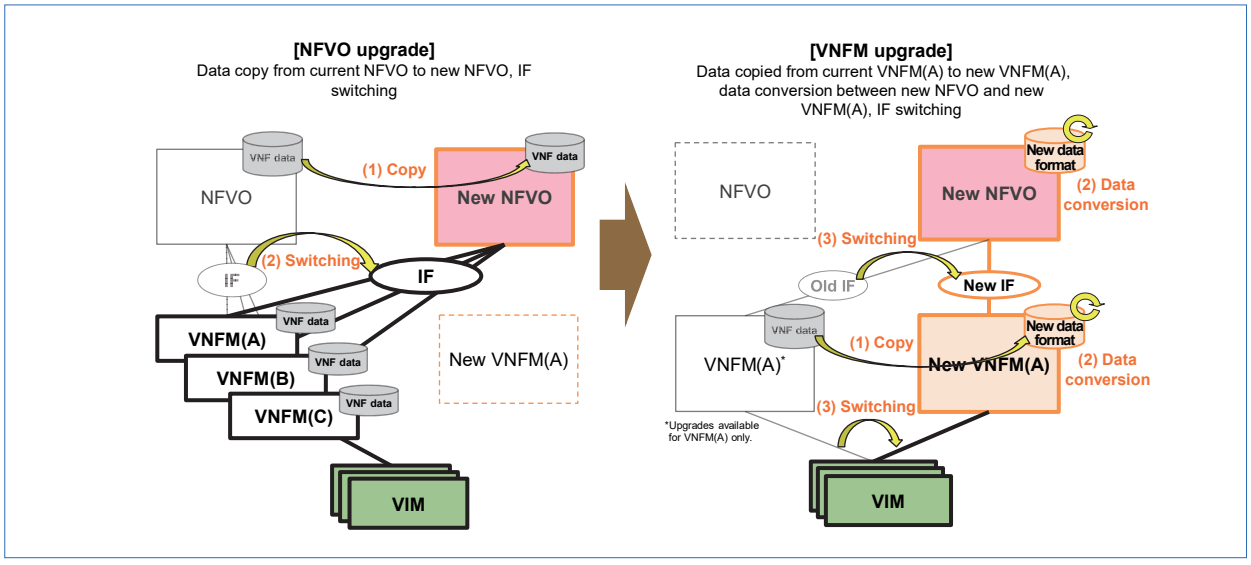


Figure 4 Upgrade method

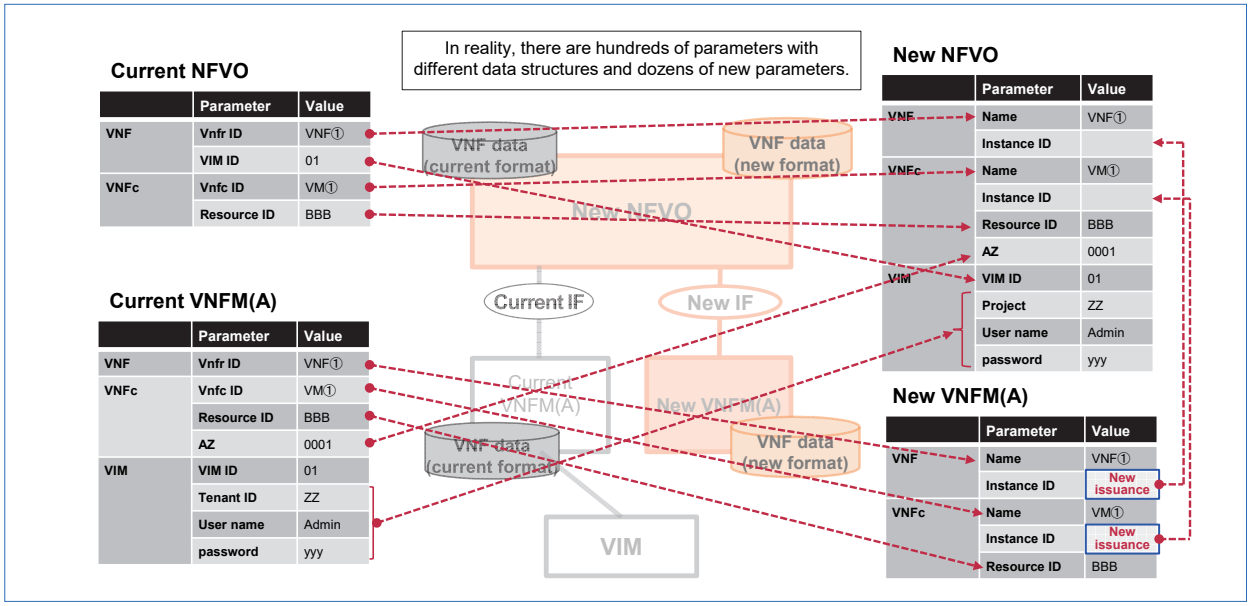


Figure 5 Data conversion and migration image

of errors in the migrated data because the data must be converted between systems manufactured by different vendors, and because various data patterns need to be considered. Therefore, as a mechanism for efficient data migration while ensuring reliability of migrated data, we considered

a method that uses the instantiation function, which is a basic lifecycle function of NFVO and VNFM, instead of developing data conversion and harmonization processes.

Instantiation is a technique to create a new VNF on NFV Infrastructure (NFVI)*37 by collaboration of NFVO, VNFM and VIM based on the VNF Package, which includes VNFD and VM images defining VNF functions and operations. The data in a new format is generated by pseudo-execution of instantiation only between new systems (NFVO, VNFM), without creating a new VNF on NFVI. This is called pseudo-instantiation (**Figure 6**). By pseudo-instantiating the VNFs that are running on the current system on a new system, VNF data in a new format are generated on the new system (NFVO, VNFM), thus enabling the migration of data of VNFs that are running without developing a new data conversion process.

To migrate data by pseudo-instantiation, the new NFVO and the new VNFM support the following functions, respectively.

(a) Automated pseudo-instantiation (new NFVO)

Conventionally, instantiation requires a maintenance technician to operate the NFVO for each VNF after registering the VNF package. However, with the migration data for more than 400 VNFs, there are concerns

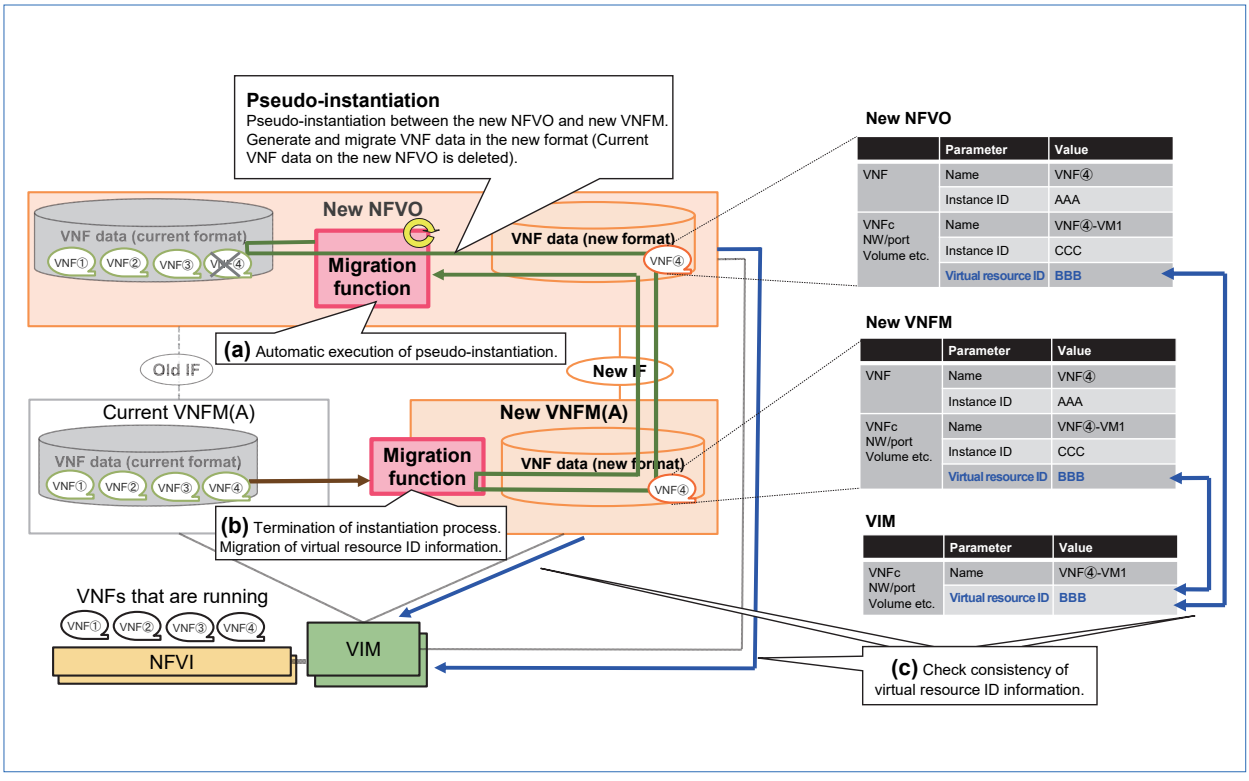


Figure 6 Mechanism of pseudo-instantiation

*37 NFVI: A generic term for the general-purpose servers, storage, and network equipment. Virtualization platforms are built using these resources.

about having to mobilize more maintenance staff and mistakes being made when making settings. To address this, the system automatically extracts the information that needs to be set during instantiation (Fig. 6 (a)).

(b) Termination of the instantiation process (new VNFM)

In normal instantiation, the VNFM performs virtual resource control to the VIM, but in pseudo-instantiation, the new VNFM terminates the process and does not perform virtual resource control to the VIM. Originally, as a result of virtual resource control to the VIM, the virtual resource ID information (ID information of virtual resources such as the generated VMs) for the deployed VNFs is collected from the VIM. However, as an alternative to this control, the virtual resource ID information is migrated from the current VNFM to the new VNFM in advance, and this information is obtained during the pseudo-instantiation process. This enables instantiation processes without implementing virtual resource control to the VIM (Fig. 6 (b)).

(c) Migration data consistency between systems (new NFVO, new VNFM)

In the new NFVO and VNFM, the virtual resource ID information generated as a result of pseudo-instantiation is checked against the virtual resource ID information managed in the VIM, and data consistency is ensured in the new NFVO, new VNFM, and VIM systems (Fig. 6 (c)).

2) Healing Continuation Mechanism

A VNF that has already migrated data through pseudo-instantiation is not able to execute its lifecycle until the switchover to the new system is completed. Also, while pseudo-instantiation improves the reliability of migrated data, it requires more time to migrate data for all VNFs since more than 400 VNFs are processed, and the period during which healing cannot be executed exceeds the allowed time (**Figure 7**).

Therefore, we designed the method to allow VNFM to perform healing alone by limiting the recovery method during the data migration period. We added a mechanism to simulate a limited period healing function that does not use the NFVO virtual resource management function (migration function) to the new VNFM, and continued healing by switching the VNFM to new NFVO connection to the VNFM to new VNFM during the data migration period (**Figure 8**). Since the NFVO virtual resource management function is not available, healing during the data migration period was limited to manual healing, which can be performed by a maintenance technician after checking the virtual resource status.

3) Reversion

To enable quick recovery and retry when a problem occurs, it is important that reversion in the upgrade process can be executed in steps that are as small as possible. In this upgrade, reversion was considered for each data migration and changeover process to the new system.

Data migration by pseudo-instantiation can be performed for each VNF selected by a maintenance

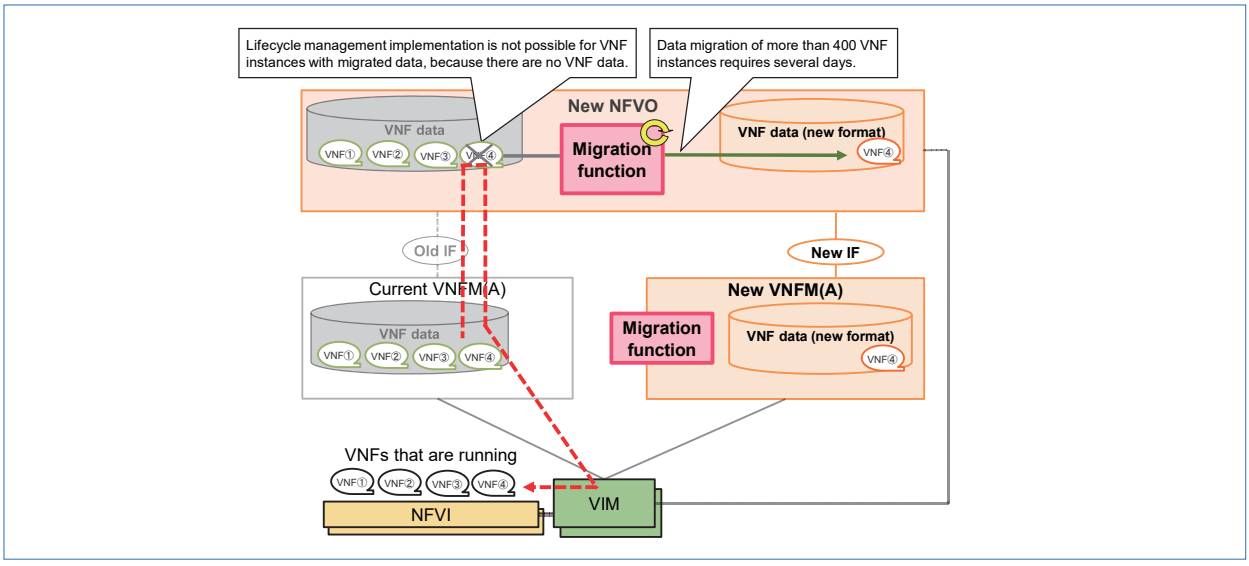


Figure 7 Challenges in continuing the VNF lifecycle during data migration

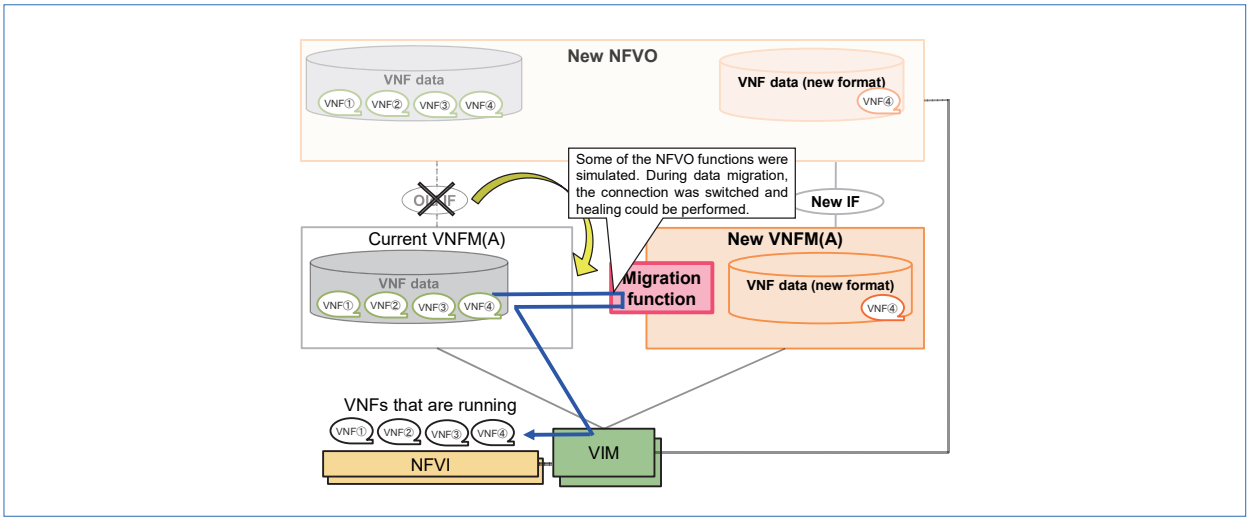


Figure 8 Mechanism for continuing healing during data migration

technician, and a rollback function is provided in case of errors. This makes it possible to revert and retry as many times as necessary, even if an issue occurs due to a VNF package creation error or data failure. In addition, when switching to the new system after data migration is completed, it is possible to switch back to the old system simply by switching networks. These were also used to deal with data correction and quick retries when problems occurred during the verification process,

which led to the completion of verification in a short period of time.

4. Conclusion

This article has described efforts to upgrade NFVO and VNFM on the NTT DOCOMO virtualization platform and a method to perform upgrades without interrupting communications services. This effort has established an upgrade method for NFVO and VNFM, as well as a multivendor environment in which ETSI NFV Stage 3 compliant products are interoperable.

To upgrade in a more efficient way, we will continue to actively contribute to ETSI NFV, collaborate with other operators in an open forum, and encourage vendors towards product commercialization. We also plan to promote and utilize Plugtest*38 [5] to ensure interoperability of multivendor products.

REFERENCES

[1] H. Oto, et al.: "Application of Virtualization Technology to Communications Networks," NTT DOCOMO Technical Journal, Vol.18, No.1, pp.4-10, Jul. 2016.
[2] T. Kamada, et al.: "Practical Implementation of Virtualization Platform in NTT DOCOMO Network," NTT DOCOMO Technical Journal, Vol.18, No.1, pp.20-28, Jul. 2016.
[3] ETSI: "Network Functions Virtualisation (NFV)." https://www.etsi.org/technologies/nfv
[4] ETSI: "Directory Listing/ISG/NFV/Open/Publications_pdf/Specs-Reports." https://docbox.etsi.org/ISG/NFV/Open/Publications_pdf/Specs-Reports
[5] ETSI: "NFV&MEC IOP Plugtests 2021." https://www.etsi.org/events/1935-nfv-mec-iop-plugtests-2021

*38 Plugtest: A joint testing program among various organizations (e.g., product vendors) performed in ETSI NFV to accelerate NFV adoption and interoperability.

CAPIF SEAL API

Standardization of Frameworks for Industrial Application Enablement in 3GPP

Core Network Development Department Yuji Suzuki

In recent years, the use of mobile networks in various industries such as automated driving, drones and smart factories has been attracting attention. 3GPP TSG SA WG6 is investigating solutions to enable a variety of industrial applications to take advantage of 3GPP networks. In particular, CAPIF, introduced in Release 15, and SEAL, introduced in Release 16, are anticipated to become service frameworks that can be commonly used in many industries. This article provides an overview of the activities of 3GPP TSG SA WG6 and describes the technical specifications of CAPIF and SEAL.

1. Introduction

Mobile network usage is expected across a wider variety of industrial fields as well as telecommunications thanks to technological advancements including, for example, the introduction of the 5th Generation mobile communications system (5G) that delivers high speed/capacity, low-latency and massive connectivity, the development of the Internet of Things (IoT) technology and the digitization of various industrial fields. In the 3rd Generation Partnership Project (3GPP), the technical specifications have been enhanced with a view to cooperating with applications outside the 3GPP domain, as exemplified by the exposure of network capabilities by the Application Programming Interfaces (APIs)*1 of the Service Capability Exposure Function (SCEF)*2 and the Network Exposure Function (NEF)*3.

3GPP TSG SA WG6 (hereinafter referred to as "SA6") has traditionally focused on solutions for mission-critical*4 communications. In addition, since

*1 API: An interface for applications to use a specific service. In this article, it refers specifically to the RESTful API.
*2 SCEF: A logical node that exposes some of the capabilities of the 3GPP system to the outside of the 3GPP domain. It is mainly used in core networks (see *13) for 4G.
*3 NEF: Similar to SCEF, a functional part of the 5G core network (see *13) that exposes some of the capabilities of the 3GPP system to the outside of the 3GPP domain.
*4 Mission-critical: A system that must be able to provide services continuously, and for which interruptions (e.g., due to failures) are unacceptable or could be extremely damaging.

Release 15, SA6 has been working on the specification of functions to support the use of 3GPP networks by applications outside the 3GPP domain to enable the aforementioned cooperation with industrial applications. For example, SA6 specified the Common API Framework (CAPIF)*5, a unified framework for northbound APIs*6 provided by 3GPP, in Release 15, and introduced Service Enabler Architecture Layer for Verticals (SEAL)*7, a set of functions that can be commonly used by various industrial applications, in Release 16.

This article presents an overview of the activities of SA6 and, in particular, describes two service frameworks, CAPIF and SEAL, which form a basis for coordinating industrial applications.

2. 3GPP SA6 Activities

Under the top-level Project Coordination Group (PCG)*8, 3GPP has three Technical Specification Groups (TSGs)*9: Radio Access Network (RAN)*10, which examines radio-related technologies; Service and System Aspects (SA)*11, which examines service functions and overall system architecture; and Core Network and Terminals (CT)*12, which examines core networks*13, terminal interfaces and functions. Each TSG is subdivided into four to six working groups. SA6 was established in 2014 as a working group within TSG SA that examines functionality and architecture, especially for mission-critical applications.

When SA6 was first launched, SA6 worked on developing specifications for mission-critical services (push-to-talk*14, video and data communications) with a focus on their use in the public safety*15 field. Later, in Release 15, the release in which the initial 5G specifications were also formulated, CAPIF was formulated as a unified framework for northbound APIs provided by 3GPP. In Release 16 and later, SA6 has developed more frameworks that can be applied to various services in addition to mission-critical services and has studied enablement of applications in various industrial fields such as Vehicle-to-Everything (V2X)*16, drones*17 and smart factories*18. Currently, the scope of SA6's work is labeled as "application enablement and critical communication applications" [1]. SA6 specifies architecture for applications that utilize the 3GPP network.

3. CAPIF

3.1 Purpose of CAPIF Implementation

CAPIF was introduced in Release 15 to provide a unified framework for the various APIs provided by 3GPP. In 3GPP, service APIs are provided by the aforementioned SCEF, NEF and functional parts for Multimedia Broadcast/Multicast Service (MBMS)*19. On the other hand, when developing and using these APIs, there are common issues to be considered, such as publishing and managing exposed APIs and security-related functions. CAPIF serves as a framework for solving these common issues.

3.2 Architecture of CAPIF

Typical CAPIF architecture is shown in **Figure 1**.

*5 CAPIF: A 3GPP framework that provides common functions for exposing northbound APIs.
*6 Northbound API: An API that is provided to a higher-level application from the perspective of the device that provides the API.
*7 SEAL: A layer of functions commonly used by multiple industrial applications using the 3GPP network.
*8 PCG: The highest decision-making body of 3GPP, responsible for overall planning and progress management of 3GPP activities.
*9 TSG: A group in 3GPP responsible for the development of technical specifications.
*10 RAN: In 3GPP, a group that is working on specifications for networks consisting of base stations, etc. that control the radio layer and are located between the core network (see *13) and the User Equipment (UE).
*11 SA: The 3GPP group handling specifications for service requirements, architectures, security, codecs and network administration.
*12 CT: In 3GPP, the group responsible for specifications for protocols within the core network (see *13) and between UE and the core network.

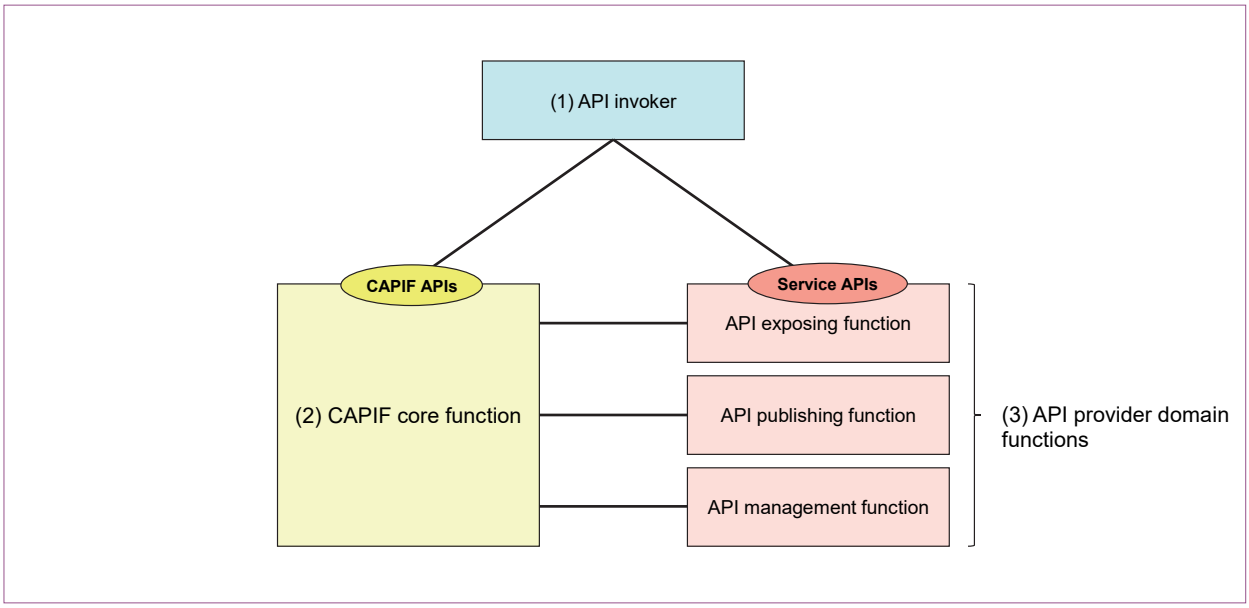


Figure 1 CAPIF architecture

1) API Invoker

API invoker invokes the CAPIF APIs and service APIs. Both applications owned by network operators and applications owned by 3rd party providers other than network operators can be API invokers.

2) CAPIF Core Function

CAPIF core function plays a central role in the various functions provided by CAPIF, such as API invoker authentication/authorization, API registration and policy management. These functions are provided as APIs, and each functional entity, including the API invoker, can use functions by invoking these APIs. The CAPIF core function is located in a domain that the mobile network operator can trust.

3) API Provider Domain Functions

The three functional entities, *API exposing function, API publishing function* and *API management function*, are collectively called API provider domain functions. CAPIF defines the provider of CAPIF core functions as the CAPIF provider and the provider of API provider domain functions as the API provider, which can be two separate entities or the same entity.

The API exposing function is a functional entity that receives service API invocations from API invokers. The API publishing function is responsible for publishing service API information to the CAPIF core function to make service APIs available to API invokers. Finally, the API management function is responsible for the management of published service APIs and has functions such

*13 Core network: A network comprised of switching equipment, subscriber information management equipment, etc. UE communicates with the core network via a radio access network.
*14 Push-to-talk: A method of voice communications in which a call can be made only while a button is pressed, as in a transceiver.
*15 Public Safety: Services for public safety, such as police, fire and emergency services.
*16 V2X: A technology that enables communication between a vehicle and its surrounding environment, such as between the vehicle and other vehicles or between the vehicle and objects on the road (such as a traffic light).
*17 Drone: An aircraft that is not piloted by a person on board. In 3GPP, systems that include drone-related functions are called Uncrewed Aerial System (UAS).
*18 Smart Factory: A factory system that utilizes IoT and other communication technologies. In 3GPP, it is specifically called Factories of the Future (FF).
*19 MBMS: A one-to-many (broadcast/multicast) communication service provided by the 3GPP system.

as auditing the service API invocation log and monitoring the service API status.

3.3 Functions of CAPIF

This section focuses on the three typical functions of CAPIF: API invoker onboarding, service API discovery and API invoker authentication and authorization. These functions are important as preliminary steps for API invokers to use a service API. For other detailed functions, please refer to [2].

1) API Invoker Onboarding

An API invoker must provide its own information to the CAPIF core function for approval prior to requesting a service API invocation. This procedure is called *onboarding*. If onboarding is successful, the API invoker will receive the information necessary for subsequent authentication and authorization.

Onboarding allows the CAPIF core function to recognize the API invoker and to authenticate and authorize the API invoker using the information obtained in this procedure. The CAPIF core function can also send information about the service APIs exposed to the API invoker.

2) Service API Discovery

For information on the service APIs that the API invoker can invoke, the API invoker can not only wait for the CAPIF core function to provide it in the onboarding procedure described above, but can also query the CAPIF core function. The procedure for obtaining service API information through this query is called service API discovery.

When an API invoker sends its identity information and the criteria for the API it wants to discover to the CAPIF core function, the CAPIF core function retrieves APIs that match the criteria from the stored API information. The CAPIF core function can also further filter the retrieved API information according to its own discovery policy. For example, it can exclude certain API categories from discovery. Sending the list of service APIs obtained in this way to the API invoker enables it to get information on target service APIs.

3) API Invoker Authentication and Authorization

An API invoker that has obtained information on a desired service API needs to go through the authentication and authorization process to invoke the service API. Authentication and authorization methods (hereinafter referred to as "security methods") used between the API invoker and the API exposing function are decided in advance between the API invoker and the CAPIF core function, based on information such as which security methods are supported by both. Three types of security methods are specified: *Pre-Shared Key Ciphersuites for Transport Layer Security (TLS-PSK)*[*20], *Public Key Infrastructure (PKI)*[*21] and *TLS with OAuth*[*22] *token* [3]. The 3GPP TSG SA WG3 (SA3) is responsible for developing these detailed security-related specifications.

After determining the security method to be used, the API invoker requests authentication and authorization from the API exposing function prior to or upon the service API invocation. The API exposing function authenticates and authorizes the

*20 TLS-PSK: A method of establishing a TLS connection that uses a Pre-Shared Key (PSK) to encrypt communications.

*21 PKI: A mechanism for certifying the registrant of a public key used in cryptography. In this article, it refers specifically to the method of establishing a TLS connection using PKI.

*22 OAuth: A standard specification for authorization of access privileges. In this article, it specifically refers to OAuth 2.0 specified in Internet Engineering Task Force Request for Comments (IETF RFC) 6749. Access privileges are controlled by issuing data called tokens.

API invoker based on a predetermined security method while cooperating with the CAPIF core function as necessary. An important role of CAPIF is to provide these security mechanisms in a unified manner.

4. SEAL

4.1 Purpose of SEAL Implementation

SEAL is a layer that brings together functions commonly used by multiple industrial applications that use 3GPP networks. In 3GPP, a vertical domain, or simply a vertical, is a set of industries and companies that provide services and products in a particular field, such as V2X, drones and smart factories. Each vertical application implements necessary functions according to its own requirements, although some functions are commonly required among multiple vertical applications. Providing such functions together as SEAL eliminates the need to implement them separately for each vertical application and leads to efficient system development.

4.2 Architecture of SEAL

Typical SEAL architecture is shown in **Figure 2**.

SEAL consists of a SEAL client, which performs client-side functions, and a SEAL server, which performs server-side functions. Both of them communicate with each other to realize functions such as location information management and group management. The SEAL server can make use of the functions provided by the 3GPP network system (e.g., APIs provided by NEF). When focusing on a particular SEAL function, a SEAL client or

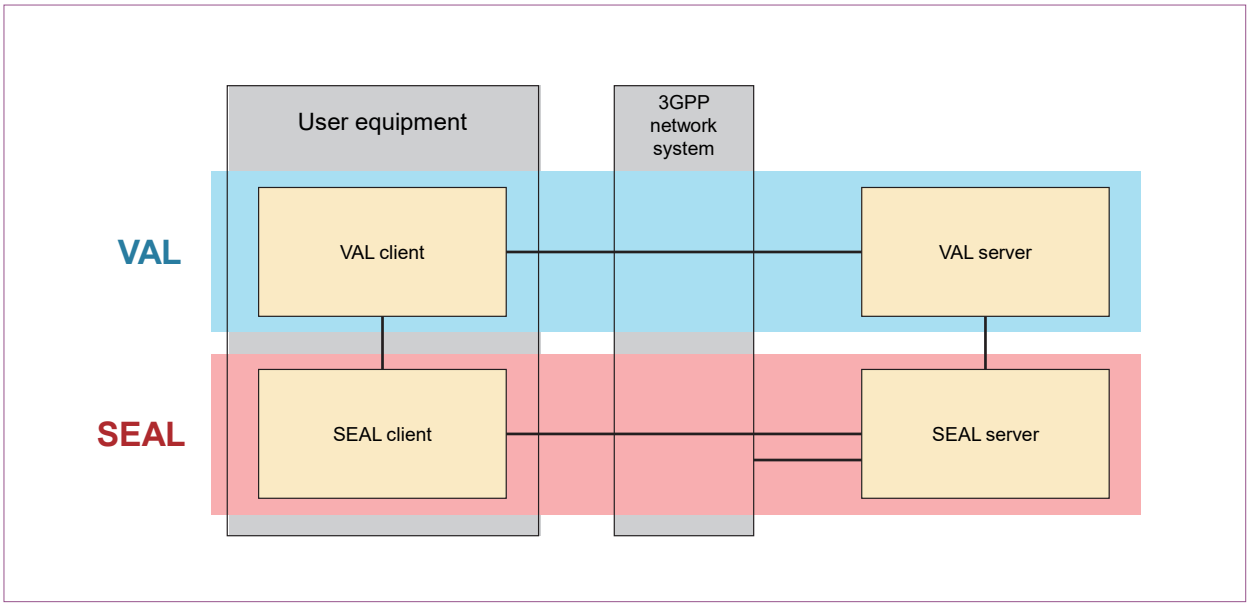


Figure 2 SEAL architecture

server may be referred to by the name of its function (e.g., a SEAL server for location management is specifically referred to as a location management server).

The Vertical Application Layer (VAL) is positioned above SEAL. VAL is the layer on which applications specific to each vertical (e.g., V2X applications) reside. VAL clients and VAL servers can use the functions provided by SEAL by communicating with SEAL clients and SEAL servers in the lower layer.

4.3 Functions of SEAL

This section focuses on three typical SEAL functions: location information management, group management and network resource management. SA6 is also studying individual solutions for each vertical and the use of the above three functions is being discussed in particular as part of these studies. For other detailed functions, please refer to [4].

1) Location Management

The SEAL client or server responsible for location management is called the location management client or server, respectively. The location management client and server can obtain information about the location of VAL service users.

For example, when the VAL server wants to obtain the location information of a specific VAL user, the VAL server sends a location reporting trigger to the location management server as a signal to start the location reporting procedure. After the location management server receives this signal, it queries the location management client for location information and sends the received location information to the VAL server.

It is possible to identify the location of other User Equipment (UE) from UE as well as from the VAL server. For example, to find out the location of UE on which location management client A is implemented, location management client B, which is implemented on different UE, can send a location reporting trigger to the location management server. As in the previous example, the location management server that receives the location reporting trigger queries location management client A for location information and sends the received location information to location management client B. This process enables the location management client to obtain the location information of the target device.

The time when location information is obtained can also be changed depending on the purpose. Specifically, when a VAL server or a location management client obtains location information of a specific VAL user, it can receive the location information immediately after sending a location information report trigger, or it can set specific conditions and receive the location information when the conditions are satisfied (e.g., the VAL server or the location management client receives the location information at regular intervals). In addition to receiving location information directly from the location management client, the location management server can also receive UE location information from the 3GPP network and send that information to the VAL server or other devices.

2) Group Management

The SEAL client and server responsible for group management are called group management client and server, respectively. The group management function allows the creation of a group consisting and multiple VAL users and management of members within that group.

For example, when a VAL user authorized to create groups wants to create a new group, the group management client of the VAL user sends a group creation request to the group management server. At this time, the identities of users to be included in the same group are also sent. Based on this request, the group management server creates a new group. After creating a group, a group management client can send a group information query request to the group management server to obtain information about the created group. The group management client can also add or delete group members by sending a group membership update request to the group management server.

It is also possible to create location-based groups when the group management service is used in combination with the location management service described above. In such a case, the group management server requests and obtains a list of users existing in a specific location from the location management server and creates a group with the users in the list. As shown in this example, multiple SEAL entities can be used to provide services.

3) Network Resource Management

The SEAL client or server responsible for network resource management is called the network resource management client or server, respectively. An example of network resource management service is to apply Quality of Service (QoS)*23 according to VAL services' requests.

To apply the desired QoS, the VAL server sends a network resource adaptation request to the network resource management server for specific UE (or UE group). Based on this request, the network resource management server allocates network resources to the target UE or UE group. The network resource management server also connects to 3GPP systems such as Policy Control Function (PCF)*24 and initiates Policy and Charging Control (PCC) procedures*25 based on requests. This makes it possible to apply QoS according to the requirements of each VAL service.

5. Conclusion

This article provided an overview of the activities of 3GPP SA6 and described the two frameworks, CAPIF and SEAL, which were standardized by SA6. These frameworks are expected to play a role in enabling industrial applications to utilize 3GPP systems. SA6 has been continuously enhancing CAPIF and SEAL since they were introduced in Release 15 and 16, respectively. NTT DOCOMO is contributing to the study of such enhancements. In Release 17 and later, the architecture for realizing edge computing*26 and technical specifications specific to services for individual industries such as V2X, drones and smart factories are also being discussed. SA6 has started technical studies for

*23 QoS: Quality of a communications service. Bandwidth and latency are typical indicators.

*24 PCF: A function of the 5G core network responsible for policy control such as QoS and billing control.

*25 PCC procedure: A series of processes related to policy control and billing control. Requests from the network resource management server are also reflected in the 3GPP system by the PCC procedure.

*26 Edge computing: A form of service in which computational processing is performed close to the UE. It is expected to reduce latency and to distribute network load.

Release 18. NTT DOCOMO will promote standardization activities in 3GPP SA6 with a view to further industrial application use cases.

REFERENCES

[1] 3GPP SP-210265: "Terms of Reference (ToR) for 3GPP TSG SA WG6 (SA6)," Mar. 2021.
[2] 3GPP TS 23.222 V17.5.0: "Functional architecture and information flows to support Common API Framework for 3GPP Northbound APIs; Stage 2," Jun. 2021.
[3] 3GPP TS 33.122 V16.3.0: "Security aspects of Common API Framework (CAPIF) for 3GPP northbound APIs," Jul. 2020.
[4] 3GPP TS 23.434 V17.3.0: "Service Enabler Architecture Layer for Verticals (SEAL); Functional architecture and information flows," Sep. 2021.

NTT DOCOMO
Technical Journal Vol.23 No.4

Editorship and Publication
NTT DOCOMO Technical Journal is a quarterly journal edited by NTT DOCOMO, INC. and published by The Telecommunications Association.

Editorial Correspondence
NTT DOCOMO Technical Journal Editorial Office
R&D Strategy Department
NTT DOCOMO, INC.
Sanno Park Tower
2-11-1, Nagata-cho, Chiyoda-ku, Tokyo 100-6150, Japan
e-mail: dtj@nttdocomo.com