Natural-language Dialogue Platform | Voice Agent | AIML

Technology Reports

# Natural-language Dialogue Platform for Development of Voice-interactive Services

*A wide variety of devices are coming to be connected to the Internet, and voice input/output is becoming an important feature for devices with no screen and for situations when user's hands are full. NTT DOCOMO has developed a natural-language dialogue platform to simplify the provision of voice-interactive services. This platform enables third-party developers to develop voice-interactive services without specialized knowledge of natural-language processing technology. In this article, we present an overview of the natural-language dialogue platform and describe several application examples.*

Service Innovation Department
*Kanako Onishi*
*Kosuke Kadono*
*Wataru Uchida* †

## 1. Introduction

Various types of voice-interactive services can now be found on the market. NTT DOCOMO, for example, provides the Shabette Concier voice-agent service that interprets the intention of the user's utterance and performs a task such as executing a mail application (task) in response to the utterance "write mail" or responding to a question like "How high is Mt. Fuji?" Going forward, we can expect such voice-interactive services to increase in number and the need for them to grow. It is therefore desirable that just about anyone be capable of developing voice-interactive services in a prompt manner.

However, there are two main issues in developing such interactive services. One is that providers and designers of services using voice interaction are not necessarily familiar with technology for achieving natural-language dialogue. The other is the need for preparing a large number of scenarios for achieving natural-language dialogue. At NTT DOCOMO, we have developed a platform*1 that enables the creation of interactive services without having to be familiar with such technology and without having to create a large number of scenarios. This natural-language dialogue platform incorporates a mechanism for using a variety of functions and external content by simply adding several lines to a set of user-system response rules (hereinafter referred to as "scenario"). It also features a mechanism for flexibly matching actual utterances with scenario data thereby reducing the cost of creating scenarios. These features make it relatively easy for developers to develop voice-interactive services.

In this article, we present an overview of NTT DOCOMO's natural-language dialogue platform and describe examples of applications that we developed using the platform.

This platform is being provided as part of the "+d*2" initiative for creating new value by sharing NTT DOCOMO business assets with partner companies.

©2016 NTT DOCOMO, INC.
Copies of articles may be reproduced only for personal, noncommercial use, provided that the name NTT DOCOMO Technical Journal, the name(s) of the author(s), the title and date of the article appear in the copies.

† Currently of Human Resources Management Department

*1 **Platform:** The basic software required to run applications.

*2 **+d:** Name of NTT DOCOMO initiative for creating new value together with partner companies.

## 2. Overview of Natural-language Dialogue Platform

Developers developing voice services and products capable of conversing with human beings can customize NTT DOCOMO's natural-language dialogue platform as needed and incorporate it in those services and products. This feature means that a desired voice agent can be easily developed by freely combining components (such as Knowledge Q&A [1]) deemed useful for developing the target voice-interactive system. The configuration of this platform is shown in **Figure 1**.

### 2.1 Function Overview

This platform controls dialogue with the user through a "scenario dialogue" function that enables a conversation consisting of a sequence of utterances and replies to be held between a person and computer and an "intention interpretation" function that classifies a user utterance into a task such as "write mail" or "check the weather." In the scenario dialogue function, the method for describing user utterances and system-response rules was designed with reference to Artificial Intelligence Markup Language (AIML)*3. Actual dialogue with the user is controlled through an AIML interpreter.

(1) AIML interpreter

The AIML interpreter analyzes the user-utterance text, references scenarios, determines the system-utterance text, and returns a reply to the user. Scenarios referenced by the AIML interpreter can be edited with management tools having a graphical user interface (GUI)*4 similar to standard text-editing tools and can be created by developers having no specialized knowledge of technology for achieving natural-language dialogue. The AIML interpreter can also extract user-related information such as interests or hobbies from user utterances based on the current scenario and can store that information in a user information database. User information extracted in this way can be used for conditional items within a scenario and incorporated in system utterances.

(2) Content collection server

The content collection server functions as a link to various types of content. It accesses external content such as weather information

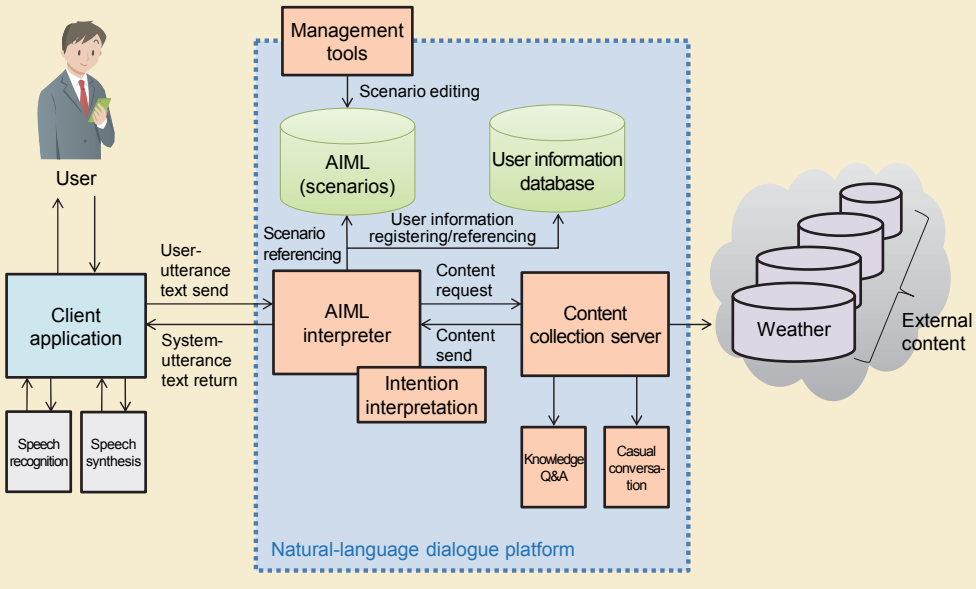


**Figure 1 Configuration of natural-language dialogue platform**

*3 **AIML:** A description technique for constructing an interactive agent.

*4 **GUI:** An interface enabling intuitive operations.

and news and collects whatever information is needed. It can also use the Knowledge Q&A function to respond to a wide range of user questions and the casual conversation [2] function to provide a reply not described in a scenario or to respond to questions that cannot be answered by Knowledge Q&A. The server sends collected information back to the AIML interpreter so that it can be used within a scenario. Instructions for collection of external content and its insertion in system responses can be described in a few lines within a scenario. This makes for a very simple design applicable to even developers with no specialized knowledge of natural-language processing technology.

## 2.2 Elemental Technologies

At NTT DOCOMO, we developed an expression-normalization function [3] for the AIML interpreter to handle diverse Japanese expressions as an elemental technology of the natural-language dialogue platform. Furthermore, for casual conversation, we developed two key functions separate from the AIML interpreter. The first of these is a user-information automatic extraction function [4] [5] for automatically extracting interests, preferences, and other user information from a dialogue. The other is a personality-oriented utterance conversion function [6] for converting a sentence written in an ordinary style to one that a certain type of personality or animated character might speak in. The above technologies make it unnecessary to describe a large number of scenarios, which means that the cost of creating scenarios can be reduced. These functions—expression-normalization, user-information automatic extraction, and personality-oriented utterance conversion—were developed at NTT DOCOMO with technical support and research results provided by NTT Media Intelligence Laboratories.

(1) Expression-normalization function

This function consolidates various types of expressions having the same meaning such as “Do you like ice cream?” and “Do you enjoy ice cream?” and “Ice cream, OK?” into a single expression. This makes it unnecessary for a developer to write up a large number of scenarios for dealing with various types of wording and enables the same effect to be obtained by simply writing a representative scenario.

(2) User-information automatic extraction function

Given a user utterance such as “I like reading,” this function would automatically extract “reading” as a user interest. In this way, there is no need for a developer to describe in a scenario a procedure for extracting user information.

(3) Personality-oriented utterance conversion function

This function automatically extracts conversion rules from previously prepared sentences having the personality, for example, of an animated character and uses such utterance-generation rules to convert a dry utterance like “Today is cold, so warm clothes are recommended.” to an utterance with more personality such as “Hey, it’s freezing outside—wear something warm!” In this way, a developer can reuse general-purpose scenarios when creating multiple characters with various personalities and can have expressions with such personalities automatically generated.

## 2.3 Dialogue Example

An example of a dialogue generated by the system that we developed using the natural-language dialogue platform is shown in **Figure 2**. In this example, the system controls the flow of a narrative conversation using the scenario dialogue function as described below.

(1) This scenario describes a sequence of actions as follows. First, the system asks the question “Where are you going tomorrow?” Next, the user responds with the name of a place. Then, on the basis of that information, the system makes a reply that includes information on a souvenir related to that place. Here, the AIML interpreter decides whether the user utterance includes a place name using a category dictionary and absorbs var-

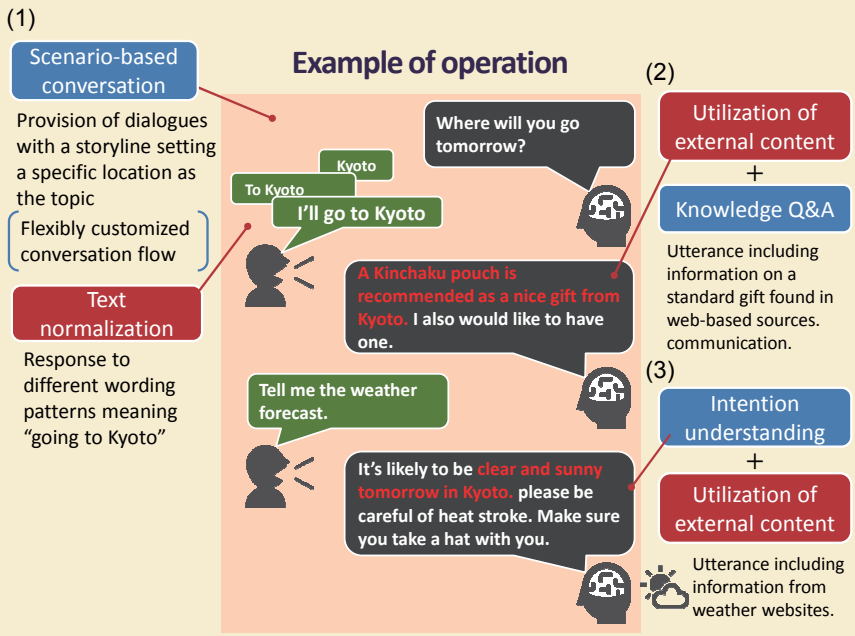


**Figure 2 Dialogue example**

iations in the user utterance using the expression-normalization function described above. In short, the system can recognize that the place where the user is going tomorrow is "Kyoto" from different wordings such as "Kyoto," "To Kyoto," and "I'm going to Kyoto" all having the same meaning.

(2) Next, given that the user's response to the system's question is a place name, the system generates a response that includes information on a souvenir related to that place using the Knowledge Q&A function. This process enables a response like "Speaking of Kyoto souvenirs, I like traditional drawstring pouches." In general, the word "Kyoto" in such an utterance would be the place recognized by the system as the one where the user will be going tomorrow and "traditional drawstring pouch" would be the name of a souvenir related to that place as obtained from the Knowledge Q&A.

(3) The final system response here is "It should be clear tomorrow in Kyoto. Be careful of heat stroke—wear a hat when going out." This is an example of a response that combines intention interpretation and external content related to the weather. The natural-language dialogue platform can generate natural conversations by combining various functions in this way.

## 3. Application Examples

The following introduces a "tablet agent," "automobile agent," and "talking toy" as application examples of NTT DOCOMO's natural-language dialogue platform.

### 3.1 Tablet Agent

1) Overview

As an application example of the natural-language dialogue platform for use in the home such as in the living room or bedroom, NTT DOCOMO has developed an agent system for demonstration purposes as a frontend application running on Android™*5 tablets. A screenshot of the tablet agent application is shown in **Figure 3**. This system enables a 3D computer-generated agent re-

*5 **Android™:** A software platform for smartphones and tablets consisting of an operating system, middleware and major applications. A trademark or registered trademark of Google Inc., in the United States.

siding on the screen to converse with the user by voice means using the natural-language dialogue platform.

2) Features

This system incorporates scenario-dialogue and casual-conversation functions and links with external content such as weather information and TV-program information. It also records user attributes and interests/preferences included in dialogues and uses that data to generate utterances that present information tailored to the individual. The emphasis here is on information useful in everyday life such as the weather, news, horoscope, and TV programming. The content collection server has the role of gathering this information.

3) Application State Transition

The application can be in one of three states: standby state, dialogue state, or notification state. A state transition occurs in the event of a user action or signal from the server. The state transition diagram is shown in **Figure 4**.

The application is in (a) standby state when no dialogue is taking place with the user or when there is no information to pass on to the user. However, on detecting an utterance of specific words or an action such as picking up the tablet (Fig. 4 (1)), the agent will move from the back of the screen to the front and enter into a state of dialogue with the user.

In the (b) dialogue state, the agent

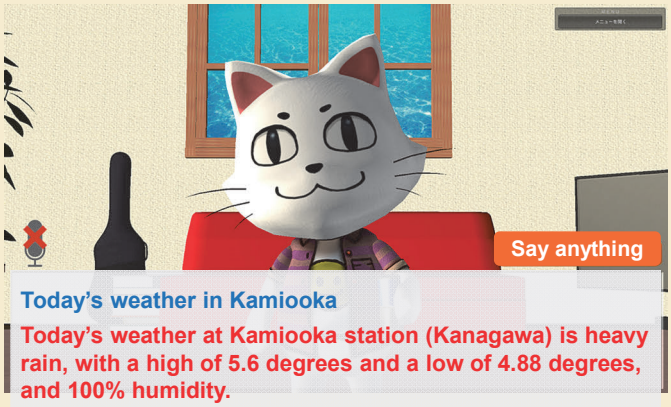


**Figure 3 Screenshot of tablet agent (dialogue state)**

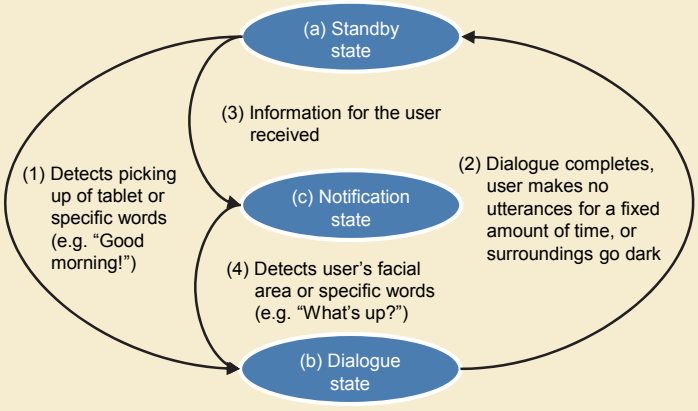


**Figure 4 State transition diagram of tablet agent**

is positioned at the front of the screen and carries on a dialogue with the user. At this time, the application sends speech data input through the microphone to the speech recognition server when not generating agent utterances. In this way, the application can always recognize user utterances. The application can also detect the user's facial area with the tablet's built-in camera and orient the agent in the direction of the user. Then, when inputting user utterances, the application can have the agent perform some kind of action, such as turning an ear toward the user, to give the dialogue a natural feel. In this state, the application can make a transition to the standby state and return the agent to the back of the screen if the dialogue completes, the user makes no utterances for a fixed amount of time, or surroundings go dark as detected by the tablet's illumination sensor (Fig. 4 (2)). Then, if information applicable to the user is received from the content collection server (Fig. 4 (3)), the application will make a transition from the standby state to the notification state.

In the (c) notification state, the application is in possession of information that needs to be passed on to the user. At this time, the agent moves to the front of the screen and performs some sort of action to get the user's attention. Then, on detecting the utterance of specific words or the user's facial area by the tablet's camera (Fig. 4 (4)), the application enters the dialogue state and presents that information to the user.

The above state-control scheme achieves an agent system that can provide support for everyday life in a natural way.

## 3.2 Automobile Agent

1) Overview

The automobile industry is one of several industries having high expectations for the industrial application of a voice agent system. Although functions for controlling on-board systems by voice have come to be provided, these functions generally accept only particular utterances such as "I'm returning home" or "I'd like to make a call." In response to this limitation, an "interactive automobile agent" was developed using the natural-language dialogue platform [7]. Assuming use by the automobile's driver or passengers, this automobile agent achieves natural-language dialogue in response to user utterances or events occurring on the automobile such as sudden braking.

2) Dialogue Example

An example of a dialogue between the user and automobile agent is shown in **Figure 5**. The interactive automobile agent detects engine startup as an event and begins to talk with the user (Fig. 5 (1)). Considering that it's the morning of a workday, the agent informs the user of weather conditions in the area of his or her workplace (Fig. 5 (2)). Then, on receiving a question about another place from the user, the agent replies based on the dialogue up to that point (Fig. 5 (3)) while also responding to an offhand utterance from the user expressing personal feelings (Fig. 5 (4)).

3) System Configuration

The system configuration of the interactive automobile agent and process flow are shown in **Figure 6**. The system links each server with a client application for use on Android smartphones and incorporates a voice-centric user interface based on a microphone and speaker mounted inside the automobile. In the system, a content-collection source such as the automobile generates event information that serves as an opportunity to initiate system-driven dialogue.

4) Features

The system links with the scenario-dialogue and casual-conversation functions and with weather information as external content. This enables the system to speak and converse with the user while including external information such as weather reports. In addition, linking with in-vehicle Controller Area Network (CAN) data*6 enables the system to generate dialogues corresponding to a variety of situations such as sudden breaking and car passing in relation to event information issued from the automobile.

## 3.3 Talking Toy

1) Overview

As an application example of the natural-language dialogue platform in

*6 **In-vehicle CAN data:** Information related to sudden braking, handling, etc. obtained using sensors mounted on the automobile.

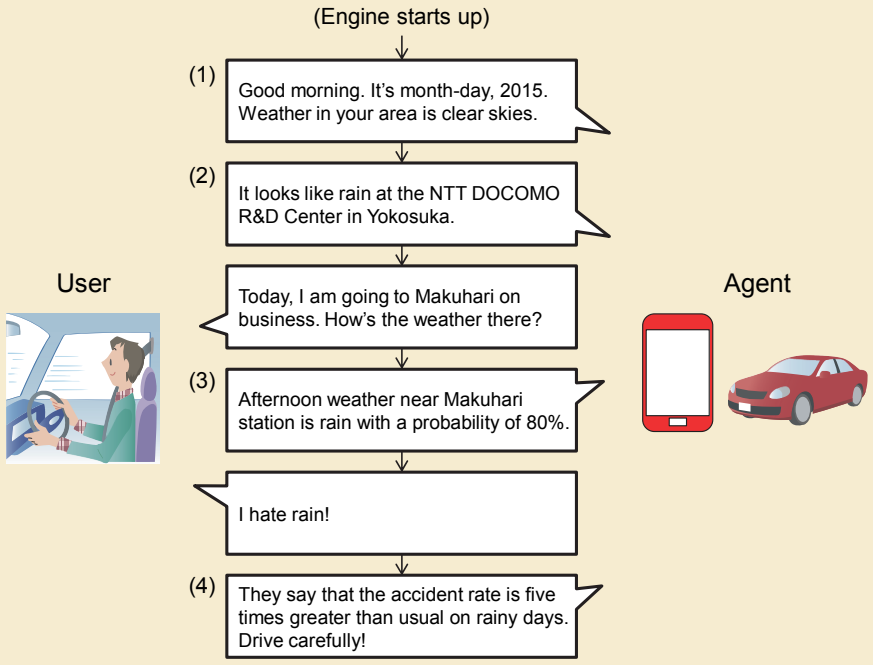


**Figure 5 Example of a dialogue with the interactive automobile agent**

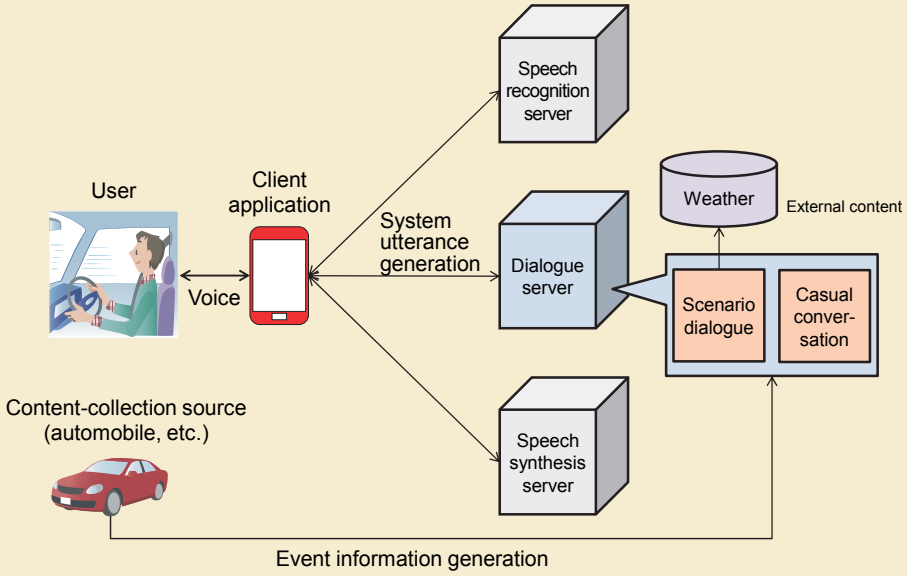


**Figure 6 System configuration of interactive automobile agent**

the toy industry targeting children and families, NTT DOCOMO and TOMY Company, Ltd. have jointly developed an interactive conversational toy named "OHaNAS®*7" that operates in conjunction with a smartphone or tablet [8].

*7 **OHaNAS®:** A registered trademark of TOMY Company, Ltd.

2) System Configuration

The system configuration of the talking toy is shown in **Figure 7**. To begin with, the user installs a specialized application in a smartphone or tablet to enable the system to connect to the natural-language dialogue platform. If the user now faces OHaNAS and makes an utterance, that speech will be connected to the terminal via Bluetooth®*8. The specialized application in the terminal then converts the speech to text using a speech recognition server and sends that text to NTT DOCOMO's natural-language dialogue platform, which returns a reply appropriate to the user's utterance. That reply is output from OHaNAS through speech synthesis. The user is able to have a natural conversation with OHaNAS in this way.

3) Features

This system links with the scenario-dialogue, casual-conversation, and Knowledge Q&A functions and with various types of external content such as weather reports and recipes. It is capable of three types of conversations between the user and toy from within the scenario dialogue function: (1) "useful conversation" in which the toy responds to the user's desire to know something, (2) "narrative conversation" in which the toy makes conversation with a storyline, and (3) "fun conversation" in which the toy becomes a playmate when the user is bored.

(1) Useful conversation

In a useful conversation, OHaNAS may reply to a user utterance such as "tell me some recipes using cabbage" or "cabbage is cheap" by saying: "I looked up some recipes using cabbage—I'll send you the information." This information would come from external content. In this example, OHaNAS provides information in response to a user request, but it may also provide the user with information directly on its own. OHaNAS does this by searching for external content based on certain rules and generating a reply even when no request is included in the last user utterance. In the above example, such a rule might be to search for and provide recipes that use an ingredient whose name was mentioned in a user utterance.

(2) Narrative conversation

A narrative conversation consists of a sequence of questions and replies that flow in a story-like manner. For example, the user may respond to the system question "Where are you going this weekend?" by saying "Nagoya," to which the system might say "Nagoya is famous for its Uiro steamed cakes." In this narrative conversation, the system proactively includes external content, which reflects the goal of achieving a toy that can be enjoyed without getting bored.

(3) Fun conversation

A fun conversation can be truly entertaining and recreational in the form of word games, riddles, quizzes, horoscopes, etc. Compared with application examples in other fields, OHaNAS includes many types of fun conversations.

Adding multiple types of conversation functions in the above way has achieved a toy with which children and families can enjoy natural-language conversation without getting bored.

## 4. Conclusion

In this article, we presented an overview of a natural-language dialogue

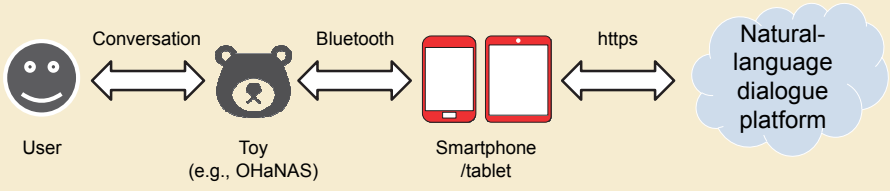


**Figure 7 System configuration of talking toy**

*8 **Bluetooth®:** A short-range wireless communication standard for interconnecting mobile terminals such as cell phones, notebook computers, and PDAs. A registered trademark of Bluetooth SIG Inc. in the United States.

platform and described three application examples. A major feature of this platform is that whatever components are needed to develop a voice-interactive system can be combined freely, which significantly improves the ability of a developer to provide a customized service. This feature makes it easy to develop distinctive voice agents in a wide range of areas including home, in-vehicle, and entertainment.

Looking to the future, we plan to apply NTT DOCOMO's natural-language dialogue platform to even more products such as home appliances and game consoles and to provide multi-language support in addition to Japanese.

## References


[1] W. Uchida et al.: "Knowledge Q&A: Direct Answers to Natural Questions," NTT DOCOMO Technical Journal, Vol.14, No.4, pp.4-9, Apr. 2013.

[2] K. Onishi et al.: "Casual Conversation Technology Achieving Natural Dialog with Computers," NTT DOCOMO Technical Journal, Vol.15, No.4, pp.16-21, Apr. 2014.

[3] T. Izumi, K. Imamura, T. Asami, K. Saito, G. Kikui and S. Sato: "Normalizing Complex Functional Expressions in Japanese Predicates: Linguistically-Directed Rule-Based Paraphrasing and Its Application," ACM Transactions on Asian Language Information Processing (TALIP), Volume 12, Issue 3, Article No.11, Aug. 2013.

[4] T. Hirano, N. Kobayashi, R. Higashinaka, T. Makino and Y. Matsuo: "Classification of Character Attributes from Self-disclosure Statements for Extracting User Information," Proc. of 21st Annual Meeting of the Association for Natural Language Processing, pp.273-276, Mar. 2015 (in Japanese).

[5] N. Kobayashi, T. Hirano, R. Higashinaka, T. Makino and Y. Matsuo: "Predicate Argument Structure Analysis of Question-Answer Pairs for User Information Extraction," 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2L3-4, 2015 (in Japanese).

[6] C. Miyazaki, T. Hirano, R. Higashinaka, T. Makino and Y. Matsuo: "Automatic conversion of sentence-end expressions for characterization of dialogue system utterances," Proc. of the 68th SIG-SLUD-B301 (Special Interest Group on Spoken Language Understanding and Dialogue Processing) of the Japanese Society for Artificial Intelligence, pp.41-46, Sep. 2013 (in Japanese).

[7] W. Uchida, "Development of Voice Agent for Communicating with Automobiles," Automotive Engineering, Vol.69, No.3, Mar. 2015 (in Japanese).

[8] NTT DOCOMO Press Releases: "DOCOMO and TOMY Develop Interactive Conversational Toy," Jun. 2015. https://www.nttdocomo.co.jp/english/info/media_center/pr/2015/0604_00.html