Technology Reports

Natural Language Processing | Voice Agent | Dialog System

# Casual Conversation Technology Achieving Natural Dialog with Computers

*Aiming to create a dialog system that will enable humans and computers to converse naturally, we have developed a casual conversation system with technological cooperation from NTT Media Intelligence Laboratories. This dialog system is characterized by its ability to correctly recognize topics and contexts of dialog, and its ability to respond in a manner similar to human beings by creating and selecting responses from large-scale data. This system holds promise of application in smart appliances or as dialog functions for domestic robots etc.*

Service & Solution Development Department

***Kanako Onishi***

***Takeshi Yoshimura***

## 1. Introduction

In recent years, voice recognition agents such as NTT DOCOMO's "Shabette Concier" have become popular. Shabette-Concier is a voice agent capable of responding to task-related utterances such as "send mail" or "call," and can answer questions such as "how high is Mount Fuji?" or "what is the highest mountain in the world?" It can also respond to casual conversation with utterances such as "I love you" or "hello." It is highly convenient for users to be able to simply make utterances to perform a task or request particular information. Nevertheless, Shabette-Concier is not only used to enjoy these conveniences – users also talk to it using a wide range of day-to-day chat, suggesting that user desire for casual conversation is very high. However, Shabette-Concier is only able to give precise replies to utterances within the bounds of assumptions, and does not have sufficient variation in its responses. Therefore, we believe we can offer casual conversation as popular content to users and expand communication module installations in new devices such as robots, games and vehicles, and apply this technology to a range of businesses to satisfy user demand for casual conversation technologies.

To respond to the user demand for casual conversation, we have developed a casual conversation system based on the technical achievements of NTT Media Intelligence Laboratories. With this system, we have aimed to enable natural conversation between computers and human beings, using utterance data created from large-scale data to generate a rich range of responses – the system does not repeat one-off utterances with users, but enables multiple and varied exchanges. This article describes an overview of the system and the dialog technology it uses.

## 2. Overview of Casual Conversation System

Here, we describe an overview of the

casual conversation system (hereinafter referred to as "Dialog System") we have developed. For "Casual conversation," we aimed to create a system that would enable dialog beginning with utterances that have no clear purpose, such as "I'd love to go to Nagano," instead of beginning with a specific request for information such as "tell me how to get to Nagano."

**Figure 1** describes an example of an application using our Dialog System. The left side shows the system utterances, while the right shows user utterances. For example, in response to the user utterance "I'm going to Nagano," the system responds with "Lucky you!" Because a general method of responding precisely and flexibly like this in casual conversation has remained unknown until now, this system is also of great significance academically.

## 2.1 System Features

This system is characterized by its ability to analyze the content of user utterances, to understand the intention or context of those utterances, and generate a response from large-scale data, rather than respond by simply matching the user utterance with preset response data. Therefore, since the system can recognize context, it's capable of natural interaction in contexts similar to human-to-human conversation.

## 2.2 System Operations

**Figure 2** describes a simplified schematic of system operations. This system basically consists of six main parts. These six parts are (1) utterance recognition section, (2) dialog control section, (3) utterance type classification section, (4) system utterance generation section, (5) system utterance selection section and (6) pre-output conversion section.

We describe processing beginning with reception of the user utterance "I want to go to Nagano."

(1) Utterance Recognition Section

The utterance recognition section analyzes the utterance received, and recognizes that the user is talking about "Nagano."

(2) Dialog Control Section

The dialog control section judges the dialog act of the user utterance, and then based on dialog history, judges whether to respond with a positive utterance, or a negative utterance etc within a broad framework. For example, if the system decides that a sympathetic utterance is required, it might respond to "I want to go to Nagano" with "what a great idea!"

(3) Utterance Type Classification Section

The utterance type classifications section generally classifies user utterances as three types - "casual conversation," "a question for the question response system" and "a question for the system itself." For example, "I want to go to Nagano" is classified as "casual conversation," "how high is Mount Fuji?" is classified as "a question for the question response system," while "what is your name?" is classified as "a question for the system itself." Thus, if the utterance is a question for the question response system, the response is made


**Figure 1 Application example using the Dialog System (left: system, right: user)**

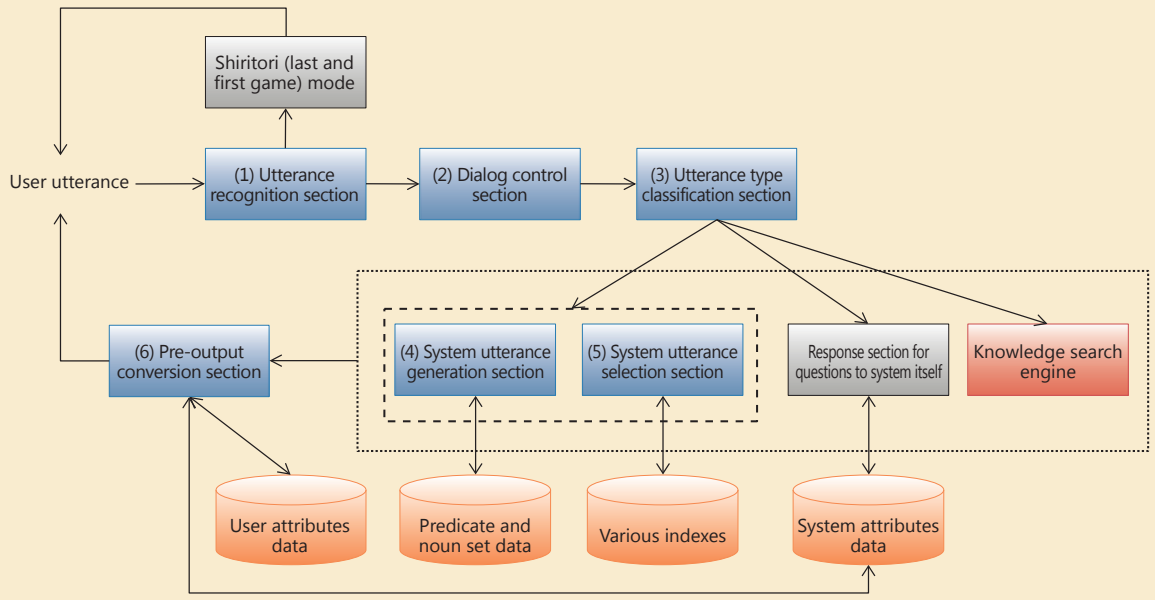


**Figure 2 System overview**

using an external knowledge search engine [1] [2]. If the utterance is a question for the system itself, the system references its attributes and responds with utterances such as "I'm 20 years old." If the user utterance is casual conversation, processing moves to the system utterance generation section to produce a response (described later).

(4) System Utterance Generation Section

The system utterance generation section uses pre-stored knowledge to generate responses. For example, for the word "Nagano," the system contains knowledge such as "go to Nagano on a school excursion," "it has many hot springs," or "the air is very fresh." Using this knowledge, the system generates utterances such as "Nagano has many hot springs, doesn't it?!" and so forth.

(5) System Utterance Selection Section

Rather than generate utterances, utterances can also be selected from large-scale utterance data. This system uses a large-scale database to store utterances created by people combined with utterances obtained from the Internet and so forth. The system selects and outputs the response it determines to be most appropriate at the present time.

(6) Pre-output Conversion Section

To perform more natural conversation, the pre-output conversion section inflects system responses. For example, the system can convert the gender-specific inflections at the end of sentences in the Japanese language to give the system a female character and a more a consistent personality.

Other than casual conversation, the Dialog System can also have modes for games etc. Modes can be switched on automatically through analysis of user utterances. Currently the game mode available is "Shiritori" ("last and first" – a word game popular in Japan).

## 3. Dialog Technology

As the most important parts of the dialogue technology in this system, we describe the focus recognition, dialog control and utterance generation sections using the example shown in **Figure 3**.

### 3.1 Focus Recognition

To hold a conversation, it's crucial that this system understands what the user is talking about. Normally, a topic (called "focus" here) can continue through a conversation, or can change with certain

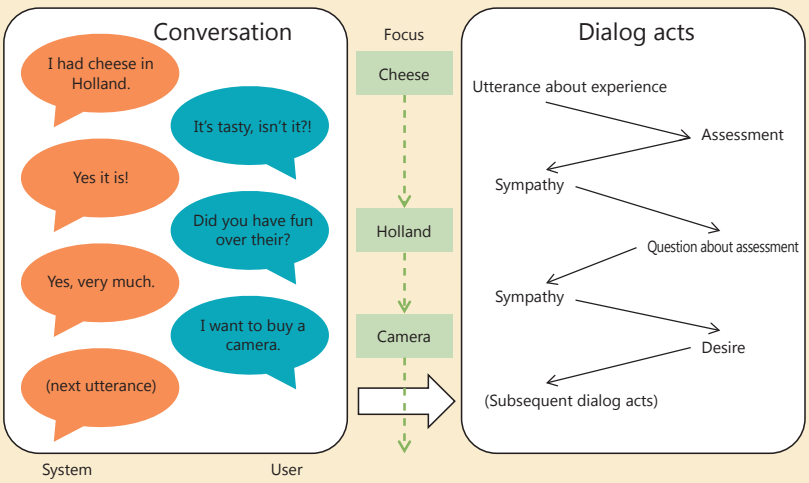


**Figure 3 Example of a conversation and its dialog act transitions**

timing. Therefore, the system performs the following two analysis processes to recognize continuance or transition of focus.

1) Extracting Focus

One analysis process extracts focus. For example, the conversation has been about Holland, up to the utterance "did you have fun over their?" and its response "yes, very much," but then a conversation about cameras begins. Here, the system has to recognize that the focus has shifted from "Holland" to "cameras." To solve this issue, the system determines and extracts appropriate vocabulary from utterances to determine focus, which is achieved by machine learning*1.

2) Anaphoric Analysis

The other analysis is anaphoric analysis*2. For example, the system says "I had cheese in Holland" to which the user responds "it's tasty, isn't it?!" In this case, the system cannot recognize "what is delicious" only by analyzing the user's most recent utterance.

To solve this issue, the system performs anaphoric analysis to complement hidden subjects or objects.

Specifically, when elements necessary for the predicate of a user utterance are missing, the system uses the current noun phrase, or the noun phrase extracted as the focus up to that point to fill in the missing elements – i.e. the utterance "it's tasty, isn't it?!" requires information about "what is tasty?" to make sense. The system estimates what can appropriately complement "what is tasty?" from the focus up to that point and noun phrases it has identified, and thus decides that "cheese" is appropriate. Pronouns such as "this" and "that" are also complemented in a similar manner from the noun phrases to which they point. For example, for the utterance "did you have fun over there?," the system estimates that "Holland" is the appropriate noun to which "over there" is referring, from the focus up to that point and noun phrases it has identified.

## 3.2 Dialog Control

Next, we describe how dialog control identifies the dialog acts of user utterances, and determines the dialog acts for utterances that should follow. Normally, people respond in conversations by thinking about the entire flow of the conversation up to the present (the context), rather than only considering the immediately previous utterance. This system considers the flow of the conversation up to the present by firstly determining the dialog acts of user utterances from several tens of classes [2]. Dialog acts include such things as "sympathy" or "assessment." Classifiers*3 to convert utterances into dialog acts are created and learn by machine learning. Data used for learning consists of large-volume utterance data with dialog acts attached.

A conversation consists of a series of dialog acts. Fig. 3 illustrates a conversation and its transitioning dialog acts. The system considers the transition of dialog acts up to the present to determine the dialog act for the subsequent utterance. To perform this, a predictor*4 is configured by machine learning. Specifically, the current user utterance dialog act, past user and system dialog acts are feature values*5, from which the system estimates likely subsequent dialog acts. For example, from the flow of "question asking an assessment," "sympathy" and "desire," the

*1 **Machine learning**: A framework that enables a computer to learn useful judgment standards through statistical processing from sample data.
*2 **Anaphoric analysis**: The process of identifying the referents for pronouns, demonstratives or abbreviated noun phrases.
*3 **Classifier**: A device that sorts inputs into predetermined groupings based on their feature values.
*4 **Predictor**: A device that estimates what will appear next from a given input.
*5 **Feature value**: Values extracted from data, and given to that data to give it features.

system might decide that "a question about a fact" should be the dialog act for the subsequent utterance, and then make an utterance such as "what kind of camera?"

In this way, updating the likelihood of subsequent dialog acts that the system might output enables the system to take on a personality. Thus, if output for "an opinion as a favorable assessment" is made to be more likely, the system would not ask a question about a fact, but would offer a favorable opinion – the system would not ask "what kind of camera?", but would make an utterance such as "cameras are fun, aren't they?!" as a favorable opinion about "cameras." This gives the system a positive character. It is also possible to set the reverse to give the system a negative character by making it more likely that it will output unfavorable opinions.

### 3.3 Utterance Generation

Here, we describe how the system generates utterances. System utterances are generated based on the dialog act to be uttered next and the current focus. Utterance generation also uses data compiled as sets of predicates and nouns that describe "what happened to what" etc. [3]. **Table 1** gives an example.

These sets of predicates and nouns are created by analyzing text on the Internet. Each one has a frequency of appearance. When this frequency is above a certain level, the text is deemed to be about matters that appear often, and therefore information commonly known by human beings. Thus, this data enables responses based on common human knowledge. For example, if the focus is "bread," the system selects the data necessary for utterance generation, as described in Table 1.

Next, declaratives*6 are created from "what happened to what" data selected to complete the utterance. For instance, if the predicate is "eat," and the object is "tasty bread," the declarative "I will eat tasty bread" is created.

Then, the most suitable declarative for the current utterance is selected from the declaratives created, and its similarity to the previous utterance is computed. In this way, whether a sentence complementing an utterance is contextually and meaningfully coherent with the flow of dialog can be checked, enabling selection of utterances that do not wildly stray from the conversation.

Finally, declaratives are adapted to particular dialog acts to convert them into utterances. For example, the sentence "I will eat tasty bread" can be converted to "I want to eat tasty bread" by the "desire" dialog act, or "you want to eat tasty bread, don't you?!" by the "sympathy" dialog act. These conversions modify declaratives in the appropriate places to give them the features required to express a particular dialog act. Using the same technology, suffixes and so forth characteristic of certain dialects can also be added, which also enables the system to express a personality.

**Table 1 Data about "what happened to what" for "bread"**

| Predicate | Noun |
|---|---|
| Eat | Tasty bread |
| Make | Bread |
| Bake | Melon bread |

## 4. Applications in Entertainment Field

In addition to the everyday conversation described, this system offers users the highly entertaining "Shiritori" (last and first) game function. As the system analyses user utterances, it will switch from conversation mode to Shiritori mode if it recognizes the command from the user to start the game. It then responds according to the rules of Shiritori. The system responds with utterances selected from vocabulary lists designed for Shiritori responses. These words are given priority based on frequency of user data, thus, the more common word, the higher the priority, and the more likely the system will utter it as a response. The game begins with commonly-known words, but gradually becomes more and more difficult with increasingly uncommon words (repeating a word is not allowed). Finally the game finishes when the user or the system can no longer respond with a word, or the game is set so that the system loses according to certain odds. When the game finishes, the system automatically switches back to conversation enabling users to continue dialog.

*6 **Declarative**: A statement that includes a subject and verb. A declarative is not a question, command or an exclamation.

## 5. Conclusion

We have described a casual conversation engine designed to enable natural dialog between humans and computers. This system is characterized by its ability to understand topics and contexts and respond to them flexibly.

The Dialog System is available through the NTT DOCOMO Innovation Village [5] or the docomo Developer support site*7. We are also developing the casual conversation functions for Drive Net Info*8. These APIs and applications will enable users to enjoy a wide range of variations with casual conversation. Into the future, we would like to expand this technology beyond mobile terminals to enable connection with other devices such as robots, games, televisions and vehicles so that users may enjoy casual conversation in a variety of scenes. This conversation technology also has potential for a variety of services. We believe that casual conversation is a technology indispensable to NTT DOCOMO as part of its mission to become a smart life partner with its customers. For example, the technology could provide companionship to people living alone or could be a partner that best understands its users – applying the Dialog System in domestic robots as well as mobile telephones etc holds the promise of infinite potential.

We plan to analyze actual usage logs to continue improving the system to enable it to hold even more human-like conversations. Furthermore, as an academic challenge, we would like to determine whether we can achieve conversation equivalent to, and indistinguishable from a real human being with this system by taking the Turing test*9 challenge.

### References


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

[2] R. Higashinaka, S. Kugatsu, W. Uchida and T. Yoshimura: "Shabette-Concier Question Response technologies" NTT GIJUTSU Journal, Vol. 25, No. 2, pp. 56-59, Feb. 2013 (in Japanese).

[3] T. Meguro, R. Higashinaka, K. Dosaka and Y. Minami: "Building a Dialog Control Unit based on analysis, and analysis of listener interaction," Information Processing Society of Japan. Vol. 53 No. 12, pp. 2787-2801, 2012 (in Japanese).

[4] H. Sugiyama, T. Meguro, R. Higashinaka and Y. Minami: "Open-domain Utterance Generation for Conversational Dialog Systems using Web-scale Dependency Structures," SIGDIAL, pp. 334-338, 2013.

[5] NTT DOCOMO: "Startup Incubation through DOCOMO Innovation Village," NTT DOCOMO Technical Journal, Vol. 15, No. 3, pp. 31-34, Jan. 2014.


*7 **docomo Developer support site**: A site that enables smartphone service and application developers to use DOCOMO APIs.

*8 **Drive Net Info**: Operated by the driver simply speaking into his or her smartphone, this is a new information service that provides information about traffic jams or the surrounding area. A trademark and registered trademark of NTT DOCOMO, INC.

*9 **Turing test**: A test of a machine's ability to exhibit intelligent behaviour, in which a human interrogator has to determine whether he or she is speaking to another a human being or a computer. If the interrogator cannot, the machine passes the test.