Grounding New Words on the Physical World in Multi-Domain Human-Robot Dialogues

Mikio Nakano[1], Naoto Iwahashi[2,3], Takayuki Nagai[4], Taisuke Sumii[2,5], Xiang Zuo[2,5]
Ryo Taguchi[2,6], Takashi Nose[2,7], Akira Mizutani[4], Tomoaki Nakamura[4]
Muhanmad Attamim[4], Hiromi Narimatsu[4], Kotaro Funakoshi[1], Yuji Hasegawa[1]
[1] Honda Research Institute Japan Co., Ltd., Wako, Saitama, Japan
[2] ATR Media Information Science Research Laboratories, Seika-cho, Kyoto, Japan
[3] National Institute of Information and Communications Technology, Seika-cho, Kyoto, Japan
[4] University of Electro-Communications, Chofu, Tokyo, Japan
[5] Kyoto Institute of Technology, Kyoto, Kyoto, Japan
[6] Nagoya Institute of Technology, Nagoya, Aichi, Japan
[7] Tokyo Institute of Technology, Yokohama, Kanagawa, Japan
nakano@jp.honda-ri.com

Abstract

This paper summarizes our ongoing project on developing an architecture for a robot that can acquire new words and their meanings while engaging in multi-domain dialogues. These two functions are crucial in making conversational service robots work in real tasks in the real world. Household robots and office robots need to be able to work in multiple task domains and they also need to engage in dialogues in multiple domains corresponding to those task domains. Lexical acquisition is necessary because speech understanding cannot be done without enough knowledge on words that are possibly spoken in the task domain. Our architecture is based on a multi-expert model in which multiple domain experts are employed and one of them is selected based on the user utterance and the situation to engage in the control of the dialogue and physical behaviors. We incorporate experts that have an ability to acquire new lexical entries and their meanings grounded on the physical world through spoken interactions. By appropriately selecting those experts, lexical acquisition in multi-domain dialogues becomes possible. An example robotic system based on this architecture that can acquire object names and location names demonstrates the viability of the architecture.

Introduction

One of the main differences between dialogues with robots and dialogues with virtual agents or telephone-based dialogue systems is that robots are in the physical world. In order for a robot to execute commands and requests by humans in the physical world, it needs to know the relationship between the linguistic expressions and physical world information obtained by sensors. For example, to respond to a request “Can you take Tom’s mug?”, it must be able to find Tom’s mug using sensors such as cameras. In this paper, we call such relationship *grounded meaning* of linguistic expressions.

Linguistic expressions and their grounded meanings used in human-robot interactions vary with houses and offices, it is not possible to prepare the knowledge on those in advance. The robot needs to acquire it through interaction with humans in a particular environment. Among the various kinds of linguistic expressions, we focus on acquiring new names in this paper. We call acquiring new words as well as their meanings *lexical knowledge acquisition*. Here, acquiring a word includes acquiring its correct pronunciation, as linguistic expressions new to robots may be out of its vocabulary.

This work is different from previous work in developmental robotics that tries to build robots that simulate child language development. We are interested in improving state-of-the-art conversational robot technologies for realistic tasks by solving a crucial problem. We therefore do not deal with general learning problems such as learning concepts of pronouns and prepositions (e.g., Gold et al. (2009)), but acquiring task-domain-specific linguistic expressions and their grounded meanings. This paper focuses on acquiring new names with grounding them on physical world using physical sensors. Note that meanings do not have to be grounded on the physical world in some task domains. For example, in the telephone directory search domain, meanings are represented by database objects. However, this paper does not deal with lexical acquisition in such domains.

Some previous work in developmental robotics that tries to enable robots (or agents) to acquire new words and their meanings from multimodal input such as a pair of speech and visual information (e.g. Roy and Pentland (2002), Yu and Ballard (2004), Iwahashi (2003)) is expected to be useful for our purpose. They, however, assume that the agent *a priori* knows that each input utterance is an instruction of a new name to the agent. In natural interaction between humans and agents, however, it is not obvious which utterance is a name instruction utterance. Usually agents, especially home and office robots, need to engage in multiple kinds of physical task domains and dialogue domains. They need to select the domain in which they should engage based on the understanding result of each human utterance. Such kind

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

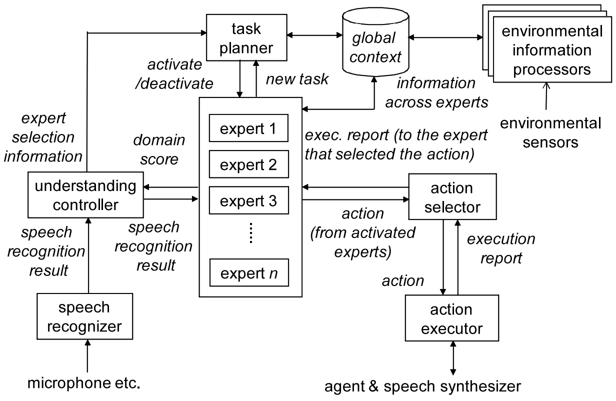

task planner
global context
environmental information processors
activate /deactivate
new task
information across experts
environmental sensors
expert selection information
domain score
expert 1
expert 2
expert 3
expert n
exec. report (to the expert that selected the action)
understanding controller
action selector
speech recognition result
speech recognition result
action (from activated experts)
execution report
action
speech recognizer
action executor
microphone etc.
agent & speech synthesizer

Figure 1: Module architecture for RIME-based systems

of interaction is called *multi-domain spoken dialogue interaction*. New name acquisition should be performed in the multi-domain spoken dialogue interactions as one of the task domains.

So far many researchers have tackled multi-domain spoken dialogue systems. They have proposed architectures for multi-domain systems (e.g., O'Neill et al. (2004), Hartikainen et al. (2004)), domain selection based on the utterance and context (e.g., Lin, Wang, and Lee (1999), Komatani et al. (2006)). None of them, however, explicitly dealt with out-of-vocabulary words and lexical knowledge acquisition.

Holzapfel, Neubig, and Waibel (2008) built a robotic system that can acquire new words using sentence patterns and its meaning. However, it does not employ multi-domain dialogue system architecture. This means that it is not easy to incorporate new task domains.

We have been developing an architecture for multi-domain conversational robots that can acquire new words and its meaning grounded on the physical world. It is based on our RIME (Robot Intelligence based on Multiple Experts) framework (Nakano et al. 2008), which is for developing the multi-domain dialogue and behavior controller for robots. RIME has modules called *experts*, each of which is specialized to perform certain kinds of tasks by engaging in dialogues and performing non-verbal actions. Our architecture employs experts for interaction for lexical knowledge acquisition as well as experts for performing other kinds of tasks. When a user utterance is detected, the robot decides which expert should be activated to deal with the utterance based on its understanding result, the context, and the situation. When an expert for lexical knowledge acquisition is selected, it tries to acquire the pronunciation of the new words through a spoken interaction. The acquired word is stored in the robot's knowledge base together with its corresponding physical world information such as image learning result and location coordinates. The acquired lexical knowledge is stored in the global context, which can be accessed from all experts, and can be used later for user utterance understanding.

Architecture for Robots That Can Acquire Lexical Information in Multi-Domain Dialogues

Multi-Expert Model

This section briefly explains the RIME framework. Its module architecture is depicted in Figure 1. It has modules called experts, each of which can engage in tasks in a single small domain. RIME also has modules for coordinating experts so that the overall system can perform multi-domain dialogue and behavior control.

Experts Each expert is a kind of object in the object-oriented programming framework. In this paper, we call tasks performed by one expert *primitive tasks*. Experts should be prepared for each primitive task type. For example, if there is an expert for a primitive task type "telling someone's phone number", "telling person A's phone number" is a primitive task. By performing a series of primitive tasks, a complicated task can be performed. For example, a museum guide robot can perform "explaining object B" by executing "moving to B" and "giving an explanation on B". Among the experts, a small number of experts can perform one or more tasks at one time. Such experts are called *activated.*

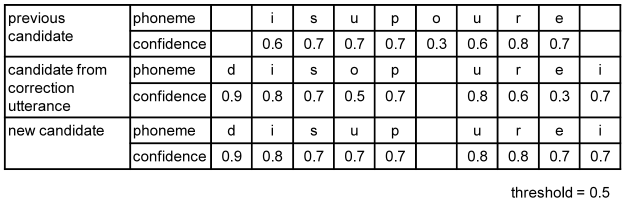
Each expert holds information on the progress of the primitive task. It includes task-type-independent information, such as which action in this primitive task is being performed and whether the previous robot action finished, and task-type-dependent information such as the user intention understanding results and dialogue history. The contents and the data structure for the task-type-dependent information for each expert can be designed by the system developer.

Interface of experts The interface of experts consists of methods for accessing its internal state. Below are some of the task-type-dependent methods, which need to be implemented by system developers.

The *understand* method updates the internal state based on the user speech recognition results, using domain-dependent sentence patterns for utterance understanding. This method returns a *domain score* which indicates the plausibility the user utterance should be dealt with by the expert. The reason why we use domain-dependent sentence patterns, which are often hand-crafted, not general parsers, is that it is not easy to build a general parser that can capture a variety of phenomena in spontaneous utterances and that those patterns are useful to estimate the domain scores.

Domain selection techniques in multi-domain spoken dialogue systems can be applied to obtain the domain score. We can employ hand-crafted rules to estimate the scores or machine-learning-based methods for estimating the score that takes into account the confidence of utterance understanding and dialogue context (Komatani et al. 2006).

The *select-action* method outputs one action based on the content of the internal state. Here, an *action* is a multimodal command which includes a text to speak and/or a physical action command. The action can be an empty action, which means doing nothing.



previous candidate	phoneme		i	s	u	p	o	u	r	e	
	confidence		0.6	0.7	0.7	0.7	0.3	0.6	0.8	0.7	
candidate from correction utterance	phoneme	d	i	s	o	p		u	r	e	i
	confidence	0.9	0.8	0.7	0.5	0.7		0.8	0.6	0.3	0.7
new candidate	phoneme	d	i	s	u	p		u	r	e	i
	confidence	0.9	0.8	0.7	0.7	0.7		0.8	0.8	0.7	0.7

threshold = 0.5

Figure 2: Example interactive word acquisition

In the definition of these methods, experts can access a common database called *global context* to store and utilize information across domains, such as information on humans, information on the environment, and past dialogue topics.

Modules coordinating experts To exploit experts, three processes, namely the *understanding controller*, the *action selector*, and the *task planner*, work in parallel.

The understanding controller receives output of an *input processor*, which typically performs speech recognition. Each time the understanding controller receives a user speech recognition result from the input processor, it performs the following process. First it dispatches the speech recognition result to all experts with their *understand* methods, which then returns the domain scores mentioned above. The expert that returns the highest domain score is selected as the expert to be activated. If the selected expert is not activated, it tells the task planner that the expert is selected be activated.

The action selector repeatedly calls the *select-action* method of each activated expert. If the returned value is not a null value, it is sent to to the action executor. Experts must return the null value when they are waiting for the finish of a robot action and when they are waiting for the user's new utterance.

The task planner is responsible for deciding which experts should be activated and which experts should not. It sometimes activates an expert by setting a primitive task, and sometimes it deactivates an expert to cancel the execution of its primitive task. To make such decisions, it receives several pieces of information from other modules. First it receives from the understanding controller information on which expert is selected to understand a new utterance. It also receives information on the finish of the primitive task from an activated expert. In addition, it receives new tasks from the experts that understand human requests. The task planner also consults the global context to access the information shared by the experts and the task planner. In this paper we do not discuss the details of task planning algorithms, but we have implemented a task planner with a simple hierarchical planning mechanism.

There can be environment sensing processes whose output is written in the global context. For example, a robot and human localization process can be used.

Experts for Lexical Knowledge Acquisition

New name acquisition in multi-domain dialogues In our architecture, some of the experts have functions for detecting new words and for lexical knowledge acquisition through spoken interaction.

When the user makes an utterance to teach a new name, one of the experts for name acquisition is supposed to be selected. Since it is not always the case that the correct phoneme sequence of the word is acquired from one utterance, it needs to request the user to confirm it. If the user acknowledges, it stores the name and its corresponding physical world information such as the image of an object and the coordinates of a location as the grounded meaning of the word. If the user makes a correction on the acquired phoneme sequence, it revises the phoneme sequence as will be explained later.

Once an expert for name acquisition is activated, it usually keeps activated until the acquisition process finishes unless the user cancels the process, because the dialogue context is considered in the expert selection. This makes it possible to understand an utterance that consists only of one new name as in the following example:

User: This is *megaqta*. (q is a double stop (short pause) in Japanese)
Robot: Is that *megara*?
User: No it's *megaqta*.
Robot: Did you say *megeqta*?
User: *Megaqta*.
Robot: *Megaqta*?
User: Yes. that's right.

In this example, the user's third utterance is just a new name. Since this is not in the vocabulary of speech recognition, it can be a recognized utterance in another domain. However, thanks to the context-dependent expert selection, the utterance is recognized as a correction to the system's confirmation request. RIME framework facilitates making this kind of interaction possible, because knowledge for interactions for word acquisition is encapsulated in each expert and just rules for estimating the domain score need to be written.

Extracting new names and estimating domain scores To extract the phoneme sequence of a new name, we use a class n-gram for name instruction utterances, where names such as object names or location names are treated as classes. In speech recognition, phoneme network is used as a language model so that phoneme sequences (phoneme recognition result) can be obtained. Words in the n-gram are tried to be matched with subsequences of the phoneme recognition result and known-word candidates are obtained with their matching score. All subsequences of the phoneme recognition result are candidates for new names. All possible sequences consisting of known words and new names are evaluated in terms of the following score and the highest scoring sequence is chosen.

$$s = w_1 s_{ngram} + w_2 \sum_i s_{matching,i}$$

Here, s_{ngram} is the n-gram score, $s_{matching,i}$ is the matching score for known word i, and w_1 and w_2 are weights. For example, let us assume the user says "korewa *megaqta* dayo (this is *megaqta*)" and it is recognized as "kareamegaqdadayo", and the sentences for training the class n-gram include “korewa ⟨object-name⟩ dayo”, where ⟨object-name⟩ denotes a class of new object names. A known word “korewa” is matched with “kare” or “karea” with high scores. Other known words are matched with subsequences in the same way. Then word sequences such as “korewa *megaqda* dayo”, “korewa *amegaqda* dayo”, and “korewa *amegaqda-dayo*” are obtained and “korewa *megaqda* dayo” is selected based on scoring.

Then the domain score is estimated based on whether the resulting word sequence matches one of the sentence patterns, the score obtained above, and the dialogue history information such as the number of turns that have been handled in the expert.

Acquiring the pronunciation of a new word through spoken interaction Although there has been a lot of work on new word acquisition, they either acquire the pronunciation from either just one utterance (Onishi, Yamamoto, and Sagisaka 2001; Bazzi and Glass 2002; Schaaf 2001; Choueiter, Seneff, and Glass 2007) or a set of utterances in off-line learning (Roy and Pentland 2002; Yu and Ballard 2004), and not many researchers have dealt with the problem with acquisition in interaction, that is, enabling the user to interactively correct the system’s pronunciation as in the example dialogue above.

On-line pronunciation acquisition from a small number of utterances is a difficult task, so some of the previous methods ask the user to spell out the name (Chung, Seneff, and Wang 2003). However, since spelling out is not effective in Japanese, we took a speech-only approach.

Our method (Sumii et al. 2010) works as follows. It first recognizes the utterance using a phoneme recognizer and matches its result to sentence patterns such as “This is ...” and “No it’s ...” to extract the phoneme sequence for the new word. At this time, the confidence score for each phoneme is also obtained. Then it matches the candidate phoneme sequence acquired so far and the phoneme sequence newly detected from the correction utterance. The matching is done by DP matching that takes into account the distance between phonemes based on the confusion matrix, and phoneme-phoneme correspondences are created. When a phoneme is correspondent with a different phoneme, the phoneme with higher confidence is chosen. If a phoneme is correspondent with the empty phoneme, it survives if its confidence score is higher than the threshold.

Figure 2 shows an example of acquiring a new word “disuprei”. If the current candidate is “isupoure” and the phoneme sequence extracted from the new utterance is “disopurei”, they are matched using DP matching, and phonemes with higher confidence are used, and the better phoneme sequence will be obtained.

Example Robotic System

This section presents an example robotic system based on the architecture described above.

Tasks and Experts

In this example, the system has the experts listed in Table 1, and can perform the following tasks:

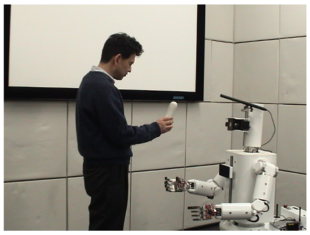


Figure 3: A snapshot of a dialogue with the implemented robotic system

- telling the name of the robot location (Expert A)
- acquiring a new location name (Expert B)
- telling the name of the object shown by a human (Expert C)
- acquiring the name and the image of a new object so that it can search for it (Experts D and E)
- finding an object whose name is specified by a human by moving (Experts F and G)
- telling someone’s phone number when requested (Experts H and I)
- executing simple commands requested by a human (Expert J)

Implementation

We implemented the above experts and combined the system with required external modules such as robot hardware, a speech recognizer, and an image processor. The robot can move using wheels and is equipped with two arms, a time-of-flight camera, two CCD cameras, a directional microphone, and three PCs for robot control, image processing and speech processing. It has a wireless network for communicating with outside computers. Dialogue behavior control and navigation are performed on outside computers.

We use Julius (Kawahara et al. 2004) as the speech recognizer together with three language models. One is a finite-state-grammar-based language model, which is the union of grammars supposed by the experts other than name acquisition experts. Each expert tries to find a recognition hypotheses that matches its own grammar among the n-best recognition hypotheses, and generates semantic representations from them. Another is a large vocabulary (about 60K words) trigram language model (Kawahara et al. 2004). Speech recognition results with this model are used for the verification of finite-state-grammar-based recognition results. The last one is a phoneme network to be used for new name acquisition.

For the image processor, we have developed a method for yielding both depth and color information in real time, by calibrating the time-of-flight and CCD cameras (Attamimi et al. to appear). Localization of the robot and people are done using ultrasonic tag sensors (Nishida et al. 2003). Figure 3 is a snapshot of a demonstration.

Currently, domain scores are estimated using hand-crafted rules that are applied to utterance verification scores,

Table 1: Experts used in the example system

expert	primitive task type	description
A	telling a location name	Understands a user request for telling the name of the area where the robot is, and looks into the database to find the name of the area. If the current robot position is not included in any area whose name has been registered, it replies "I don't know the name of this location."
B	acquiring a location name	Acquires the pronunciation of a location. When pronunciation acquisition finished, it stores the location information in the form of a pair of location coordinates and the acquired pronunciation.
C	telling an object name	Understands a user request for telling the name of an object and communicates with the object image recognizer to get the ID of the object that is being shown, then it looks into the database to find the name of the object. If it cannot get the object ID from the image recognizer, it replies "I don't know the name of the object."
D	acquiring an object name	Acquires the name of an object by extracting a new phoneme sequence from a name instruction utterances
E	learning an object image	Asks the user to hand over the object to learn, then learns its image while rotating the object. When the image learning finishes, it stores the acquired name of the object together with the ID of its image.
F	understanding a request for searching for an object	Understands human requests to search for an object and telling the name of an object to somebody through a dialogue.
G	searching for an object	Makes the robot move around while the image recognizer is searching for the object. When it finds something similar to the object using a color histogram and depth information, it makes the robot get closer to it and recognizes the image using SIFT features. When it finds the object, it tells the user it has found it.
H	understanding a request for phone number	Understands human requests for a phone number of a specified person. After performing some dialogue management, when it finishes understanding, it tells the task planner the new task to tell the phone number.
I	telling a phone number	Searches for the requested phone number in the database and tells it to the user.
J	reacting to a user command	Understands a simple command by a human and selects an action using a set of command-action rules.

dialogue history, and the scores obtained in the new word extraction process, although we think machine learning techniques can be used for estimating the domain score.

Figure 4 is an example interaction between the robot and a human user that demonstrates current implementation status. Note that the interaction was done in Japanese, but only the translations are written in the figure.

Conclusion and Future Work

This paper described our effort for developing an architecture for robots that can acquire new words and their meanings while engaging in multi-domain dialogues. The implementation of an example system has suggested the proposed architecture and its underlying RIME framework is viable.

Among many pieces of work yet to be done, the following issues are worth mentioning. First, we plan to evaluate the domain selection and interactive lexical acquisition in detail. Second, the current domain selection is based only on speech. We have already developed a method for detecting utterances directed to the robot by differentiating them from utterances directed to other humans (Zuo et al. 2010), and we are planning to incorporate it into our architecture. Third, although we have dealt only with name instruction utterances, we will develop a method for detecting out-of-vocabulary words in other types of utterances. For example, if the robot can detect an out-of-vocabulary word in utterances such as "can you search for *megaqta*?", it can invoke the lexical acquisition process. We are currently working on making this type of interaction possible. Finally, we plan to expand the area we tackle from just lexical acquisition to the acquisition of the domain-dependent grammatical knowledge for understanding utterances that are not covered by the pre-defined grammar.

References

Attamimi, M.; Mizutani, A.; Nakamura, T.; Nagai, T.; Funakoshi, K.; and Nakano, M. (to appear). Real-time 3D visual sensor for robust object recognition. In *Proc. IROS-2010*.

Bazzi, I., and Glass, J. R. 2002. A multi-class approach for modelling out-of-vocabulary words. In *Proc. 7th ICSLP*, 1613–1616.

Choueiter, G. F.; Seneff, S.; and Glass, J. R. 2007. New word acquisition using subword modeling. In *Proc. Interspeech-2007*, 1765–1768.

Chung, G.; Seneff, S.; and Wang, C. 2003. Automatic acquisition of names using speak and spell mode in spoken dialogue systems. In *Proc. NAACL-03*, 32–39.

Gold, K.; Doniec, M.; Crick, C.; and Scassellati, B. 2009. Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence* 173:145–166.

Hartikainen, M.; Turunen, M.; Hakulinen, J.; Salonen, E.-P.; and Funk, J. A. 2004. Flexible dialogue management using distributed and dynamic dialogue control. In *Proc. Interspeech-2004 (ICSLP)*, 197–200.

Holzapfel, H.; Neubig, D.; and Waibel, A. 2008. A dialogue approach to learning object descriptions and semantic categories. *Robotics and Autonomous Systems* 56(11):1004–1013.

Agent	utterance and behavior	the system's internal process
User	Hi Robot!	Expert J is selected to understand the utterance.
Robot	How may I help you?	Expert J selects the system utterance using a command-action rule.
User	Do you know the name of this location?	Expert A is selected.
Robot	Sorry I don't know the name of this location. (shake the head)	Expert A searches for the name of the area in the databases and tells the user that it does not know the name.
User	The name of this location is *robonoma*.	Expert B is selected and a new location name is extracted.
Robot	*roboroma?*	Expert B requests confirmation of the new name
User	*robonoma*.	Expert B continues to be selected based on the context. It updates the name using the interactive new name acquisition method.
Robot	Is it *romomoma?*	
User	*Robonoma*.	
Robot	Is it *bobomoma?*	
User	*Robonoma*	
Robot	Is it *robonoma?*	
User	Yes.	Expert B understands the acknowledgment and stores the a pair of the location coordinates and the acquired name.
Robot	Okay.	Expert B tells the user that it successfully acquired the name and finishes the primitive task.
User	By the way, do you know Mr. Torii's phone number?	Expert H is selected.
Robot	Mr. Torii's phone number?	Expert H requests confirmation since the confidence score of the user utterance result is not high enough.
User	Um, forget it.	Expert H is selected to understand the user's utterance to cancel the task.
Robot	Okay.	Expert H tells the user that it successfully canceled the task.
User	Do you know the name of this?	Expert C is selected.
Robot	Sorry I don't know its name. (shake the head)	Expert C communicates with the image processor to get the ID of the object image and finds that it is a new object. It tells the user that it does not know the object.
User	This is *migurikon*. (Japanese: kore no name wa migurikon dayo)	Expert D is selected and extracts a new name candidate, but it fails to identify the part of the new name in the utterance. It obtained the word sequence "kore wa *namayamigurikon* dayo" (English: this is *namayamigurikon*)
Robot	Is it *namayamigurikon*?	Expert D requests confirmation.
User	*Migurikon*.	Expert D continues to be selected and updates the new word candidate.
Robot	Is it *migurikon*?	Expert D requests confirmation.
User	Yes.	Expert D understands the acknowledgment, writes the new word into the global context, and finishes the primitive task.
Robot	I'll remember it. Can I see it? (reach out the right hand)	Expert E is activated by the task planner. It will learn the image of the object and store the new name and image ID.

Figure 4: Example interaction

Iwahashi, N. 2003. Language acquisition through a human-robot interface by combining speech, visual, and behavioral information. *Information Sciences* 156:109–121.

Kawahara, T.; Lee, A.; Takeda, K.; Itou, K.; and Shikano, K. 2004. Recent progress of open-source LVCSR engine Julius and Japanese model repository. In *Proc. Interspeech-2004 (ICSLP)*, 3069–3072.

Komatani, K.; Kanda, N.; Nakano, M.; Nakadai, K.; Tsujino, H.; Ogata, T.; and Okuno, H. G. 2006. Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors. In *Proc. 7th SIGdial Workshop*, 9–17.

Lin, B.; Wang, H.; and Lee, L. 1999. Consistent dialogue across concurrent topics based on an expert system model. In *Proc. 6th Eurospeech*, 1427–1430.

Nakano, M.; Funakoshi, K.; Hasegawa, Y.; and Tsujino, H. 2008. A framework for building conversational agents based on a multi-expert model. In *Proc. 9th SIGdial Workshop*, 88–91.

Nishida, Y.; Aizawa, H.; Hori, T.; Hoffman, N.; Kanade, T.; and Kakikura, M. 2003. 3D ultrasonic tagging system for observing human activity. In *Proc. IROS-2003*, 785–791.

O'Neill, I.; Hanna, P.; Liu, X.; and McTear, M. 2004. Cross domain dialogue modelling: an object-based approach. In *Proc. Interspeech-2004 (ICSLP)*, 205–208.

Onishi, S.; Yamamoto, H.; and Sagisaka, Y. 2001. Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes. In *Proc. Eurospeech-2001*, 693–696.

Roy, D., and Pentland, A. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science* 26:113–146.

Schaaf, T. 2001. Detection of OOV words using generalized word models and a semantic class language model. In *Proc. Eurospeech-2001*, 2581–2584.

Sumii, T.; Iwahashi, N.; Funakoshi, K.; Nakano, M.; and Oka, N. 2010. A speech interface for correcting misrecognized phonemes in out-of-vocabulary word acquisition. In *Proc. Annual Conference of Japanese Society for Artificial Intelligence*. (in Japanese).

Yu, C., and Ballard, D. 2004. On the integration of grounding language and learning objects. In *Proc. 19th AAAI*, 488–494.

Zuo, X.; Iwahashi, N.; Taguchi, R.; Matsuda, S.; Sugiura, K.; Funakoshi, K.; Nakano, M.; and Oka, N. 2010. Robot-directed speech detection using multimodal semantic confidence based on speech, image, and motion. In *Proc. ICASSP-2010*, 2458–2461.