Development of an Interactive Navigation System based on Household Ontology and Commonsense reasoning

Alan Schalkwijk
Graduate School of Science and Engineering
Aoyama Gakuin University
Sagamihara Kanagawa Japan
c5620170@aoyama.jp

Motoki Yatsu
College of Science and Engineering
Aoyama Gakuin University
Sagamihara Kanagawa Japan
yatsu@it.aoyama.ac.jp

Takeshi Morita
College of Science and Engineering
Aoyama Gakuin University
Sagamihara Kanagawa Japan
morita@it.aoyama.ac.jp

ABSTRACT

In recent years, researchers from the fields of computer vision, language, graphics, and robotics have tackled Embodied AI research. Embodied AI can learn through interaction with the real world and virtual environments and can perform various tasks in virtual environments using virtual robots. However, many of these are one-way tasks in which the interaction is interrupted only by answering questions or requests to the user. In this study, we aim to develop a task-oriented interactive system using virtual household ontology and commonsense reasoning, in which a virtual robot can reason about the location of a guide while interacting with the user and guide the user around the house.

CCS CONCEPTS

• Information systems → Information systems applications.

KEYWORDS

Navigation System, Ontology, Dialog System, Commonsense Reasoning

1 Introduction

In recent years, research has been conducted on Embodied AI [8], which attempts to solve tasks such as object navigation and question answering in virtual spaces by using agents such as virtual robots that learn in the real world and virtual spaces to guide the user to the desired object or room. In order to solve these tasks, it is necessary to integrate technologies from various fields such as computer vision, computer graphics, natural language processing, artificial intelligence, and robotics.

Traditional research in computer vision and natural language processing has been developed by 'Internet AI', which focuses on pattern recognition of images and text. Embodied AI focuses on the ability of a virtual robot to see, hear, speak, move, etc. in a virtual space.

In conventional Embodied AI research, tasks such as object navigation and question answering can be solved by a one-way dialogue in which the virtual robot simply responds to specific requests from the user. In order for the virtual robot to be able to respond to ambiguous requests from the user, it is necessary for the virtual robot to interact multiple times with the user and infer his intentions.

In this research, we aim to develop a two-way interactive navigation system by introducing knowledge inference techniques to Embodied AI research. Specifically, based on the home environment ontology and common-sense reasoning, we attempt to solve a task in which a virtual robot can reason about the user's intentions while interacting with the user multiple times, and guide the user to the location in the virtual home environment that the user needs.

2 Related research

2.1 Vision and Dialog Navigation

Vision and Dialog Navigation (VDN)[1] is a task in which a human gives unspecified instructions to a virtual robot, such as in a home environment, and the robot responds and guides the human. The VDN aims at learning a virtual robot that can automatically guide a human in an unfamiliar place, such as a verbal tele-operated home robot or an office robot, which would await the user to give clues for predicting the facility in need." The VDNs are trained and simulated using a realistic simulation environment called Matterport Room2Room[11], by creating a dataset of dialogue examples and the virtual robot's actions (forward and turn) in response to the dialogue examples.

The main dialogue examples in this task are ambiguous ones such as "Go to the room with the bed" rather than direct actions such as "Go right". In this study, we use VDNs to capture the images of the virtual robot and the ambiguous speech of the user to guide the robot to the location. We aim to estimate and guide the user to the location based on more ambiguous and everyday speech such as "I'm hungry" than the dialogue examples assumed in the VDN.

2.2 Research on Common Sense

Previous researches on common sense knowledge include COMET-ATOMIC [2], a commonsense knowledge graph with 1.33 million English entities and tuples of everyday reasoning knowledge about events.

COMET-ATOMIC enables us to reason about everyday events and objects, as well as human behavior and mental states in relation to certain events, based on 23 kinds of common-sense relations. Using this knowledge graph, we have also created a language model that performs inference based on two inputs: entity pointing at an event or a thing, and predicate denoting a relation.

In this research, we use this learned language model to infer the location of a guide from the user's speech. In this study, we use three types of common-sense relations: "events that humans cause in response to certain events," "actions that humans need to take in response to certain events," and "places where certain objects may be found".

3 Proposed System

In this section, we describe the structure of the proposed system, the construction procedure of the home environment ontology, and the dialogue based on common sense reasoning.

In the proposed system, a user inputs a sentence to the system, and a virtual robot responds and guides the user through the virtual home environment based on the dialogue rules. We assume that the user and the virtual robot are walking together in the virtual home environment, and that they are looking at the same scenery (images of the virtual environment sensed by the virtual robot's camera).

3.1 System configuration

Figure 1 shows the system configuration for this research. First, the conversation between the user and the system is conducted in text format through the interface. When a user inputs an utterance, the intention and entity of the utterance are extracted through the natural language understanding module in the dialogue system. The dialogue management module then determines the system's response based on the intentions, entities, and dialogue rules. When a virtual robot is to guide the user based on the dialogue rules, an action script is specified to move the robot, and the action is executed on the Embodied AI simulator based on the script. In case the user's speech is ambiguous and it is difficult to specify the location of the guidance, we use a reasoning module that utilizes the home environment ontology and common-sense reasoning to infer the location of the guidance from the speech.

3.2 Interface

In this study, we use Telegram[1] as an interface, which is an instant messaging application that allows users to talk by text input. We chose Telegram because it is easy to install and widely used, and because it can be linked with Rasa [5], the dialogue framework used in this study.

3.3 Dialogue System

To build this dialogue system, we use Rasa [5], an open-source dialogue framework designed to develop task-oriented dialogue systems.

As shown in Fig. 2, Rasa prepares a data set of example dialogues corresponding to certain intentions created by the user, and by learning from the data set attached with prediction labels, it is able to extract intentions and entities from the user's input. Then, dialogue is performed based on dialogue rules. In the dialogue rules, the system can determine the output content of the system based on the intentions extracted from the user's input. Furthermore, specific programs written in Python can be executed based on the dialogue rules.

Next, we discuss our reasons for selecting Rasa. We surveyed the latest papers on dialogue system frameworks [7, 9] and compared Rasa with other frameworks. As a result, we found that Rasa is superior in the following points:

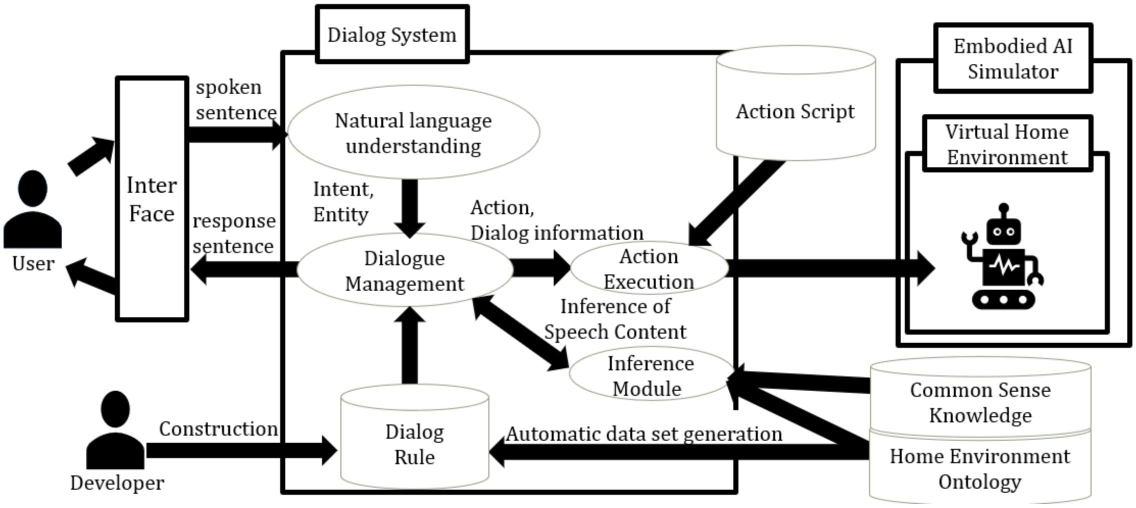
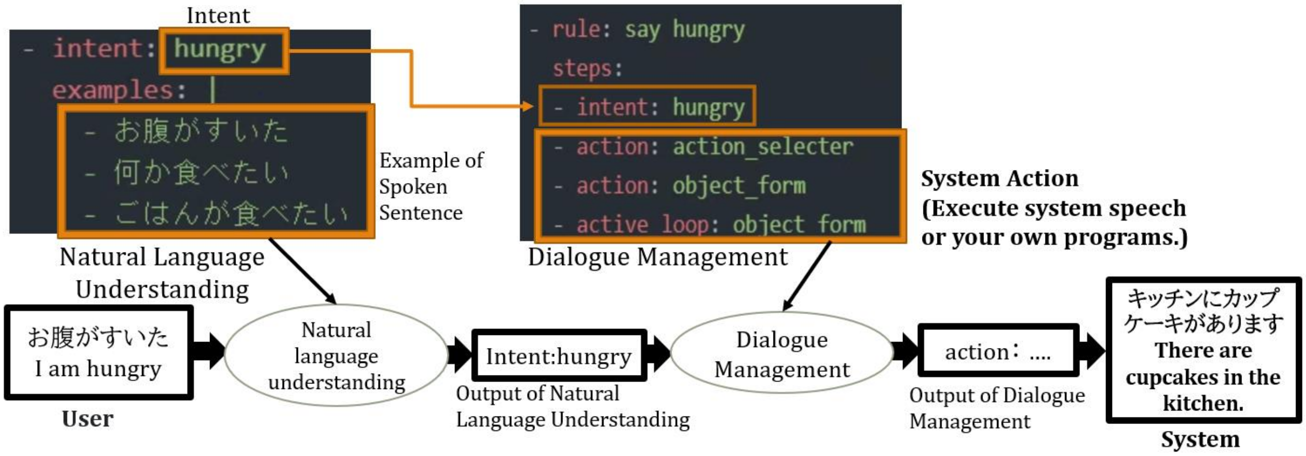


Figure 1: System configuration diagram

[1] https://telegram.org



```
- intent: hungry
  examples: |
    - お腹がすいた
    - 何か食べたい
    - ごはんが食べたい
```

```
- rule: say hungry
  steps:
  - intent: hungry
  - action: action_selecter
  - action: object_form
  - active_loop: object_form
```

Intent
Example of Spoken Sentence
Natural Language Understanding
Dialogue Management
System Action (Execute system speech or your own programs.)
お腹がすいた I am hungry
User
Natural language understanding
Intent:hungry
Output of Natural Language Understanding
Dialogue Management
action：....
Output of Dialogue Management
キッチンにカップケーキがあります There are cupcakes in the kitchen.
System

Figure 2: Example of system input/output

- It can extract intentions and entities from user input sentences.
- Ability to configure dialogue rules and slots
- Capable of executing programs created by the user based on dialogue rules
- Can be integrated with existing interfaces
- The development scale is large, and the framework is well documented and easy to use.

For these reasons, we decided to use Rasa for the development of the dialogue system in this study.

3.4 Embodied AI Simulator

We use VirtualHome [6] for the Embodied AI simulator, which is a simulator that aims to simulate activities in a virtual home. The environment of VirtualHome is represented as a knowledge graph, which contains environment data such as room and object identification numbers, location coordinates, and object states, in the form of a dictionary with data names as keys and corresponding values. In this research, we use this environmental knowledge to construct a home environment ontology.

In VirtualHome, specific tasks can be performed using action scripts. As shown in Fig. 3, an action script consists of the actions of the virtual robot, the rooms and objects to be acted upon, and the identification numbers of the rooms and objects. In this research, this action script is used to guide the robot.

We compared VirtualHome with other Embodied AI simulators, referring to a survey paper on the latest Embodied AI simulators [8]. As a result, we found that VirtualHome has rich environment data about objects in the environment. As a result, we decided to use VirtualHome because it is superior in the following three aspects: it has rich environment data about objects in the environment, it is easy for virtual robots to move to specific rooms and objects, and it provides a virtual home environment.

$$[\text{Walk}]\ \langle \text{living\_room} \rangle\ (1)$$

Figure 3: Example of action script

3.5 Home environment ontology

There is prior work on ontologies for behavior in VirtualHome[13].In this study, we created a new ontology because we need information such as properties and Japanese labels for objects and rooms that exist in VirtualHome.

Figure 4 shows the construction procedure of the home environment ontology. In Step 1 of Figure 4, we extract the names of rooms and objects from all environment data in the VirtualHome knowledge graph. In Step 2, we add the extracted data as classes of the ontology using DODDLE-OWL[3] and WordNet[4]. In Step 3, Japanese labels are added to the added classes in order to perform Japanese dialogues. Next, in Step 4, all environment data is generated as an instance of the added class. In Step 5, we define properties to store environment data such as location coordinates, identification number, and status in the created instance. In addition, to facilitate guidance, we added a property to indicate in which room an object exists. In the end, the number of classes, properties, and instances of the home environment ontology created was 239, 6, and 356, respectively. Figure 5 shows part of the class hierarchy of the created ontology, and Table 1 shows the properties and their definition and value ranges. This ontology is written in RDF/XML format.

3.6 Inference of Speech Content

Figure 6 shows the flow from user speech to location guidance based on the home environment ontology and common-sense reasoning.When a user inputs an ambiguous utterance such as "I am Hungry" in Japanese. The system translates it into English as "I am hungry" in order to perform inference using COMET-ATOMIC[2]. Next, the translated English sentence is input into COMET-ATOMIC, and the system uses the three common sense

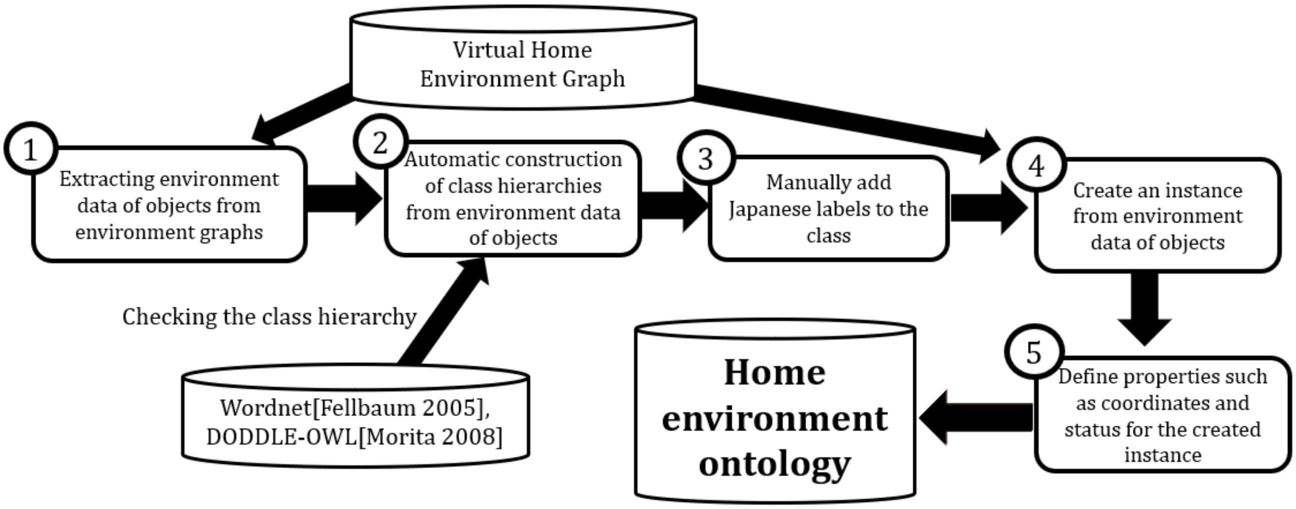

Virtual Home Environment Graph
1
Extracting environment data of objects from environment graphs
2
Automatic construction of class hierarchies from environment data of objects
3
Manually add Japanese labels to the class
4
Create an instance from environment data of objects
5
Define properties such as coordinates and status for the created instance
Checking the class hierarchy
Wordnet[Fellbaum 2005], DODDLE-OWL[Morita 2008]
Home environment ontology

Figure 4: Home environment knowledge construction procedure

Property	Domain	Range
https://schema.org/identifier (object ID)	Matter,Object,Room	xsd:int
https://schema.org/longitude (x coordinate)	Matter,Object,Room	xsd:double
https://schema.org/latitude (y coordinate)	Matter,Object,Room	xsd:double
https://schema.org/elevation (z coordinate)	Matter,Object,Room	xsd:double
https://schema.org/object (object state)	Object	xsd:string
https://schema.org/containedInPlace (location of the room with the object)	Matter,Object	Room

Table1: Properties and their domains and ranges

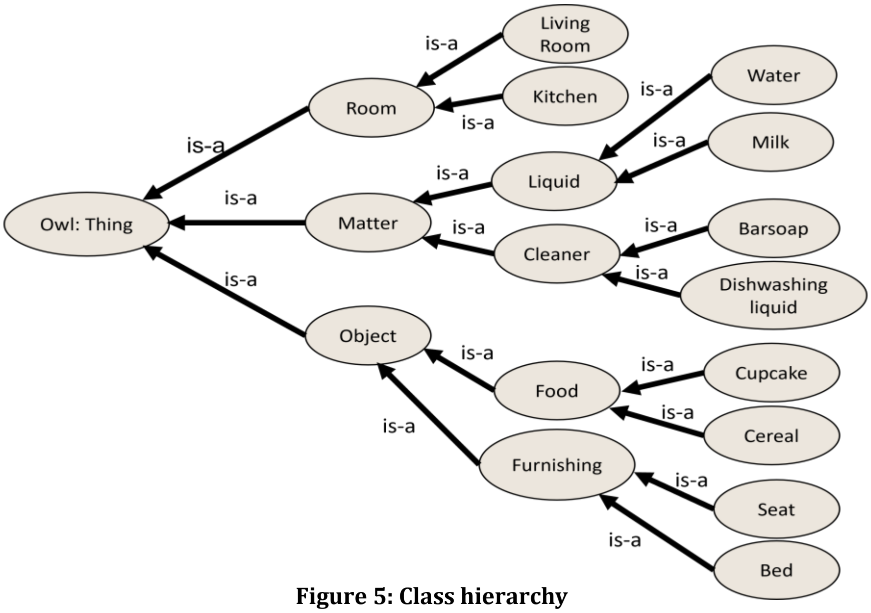


Figure 5: Class hierarchy

relations of "events that humans cause in response to certain events," "actions that humans need to take in response to certain events," and "places where certain objects may be found" to infer where the user wants to go and what actions to take. As a result, output results such as "go to kitchen" and "eat food" are obtained. The system then extracts nouns from the output and obtains the names of places the user wants to go and things the user wants, such as "kitchen" and "food". The obtained names are checked against the classes and subclasses in the home environment ontology. The system obtains the property values and Japanese labels for the location of the room from the instances of the matched class, and presents the user

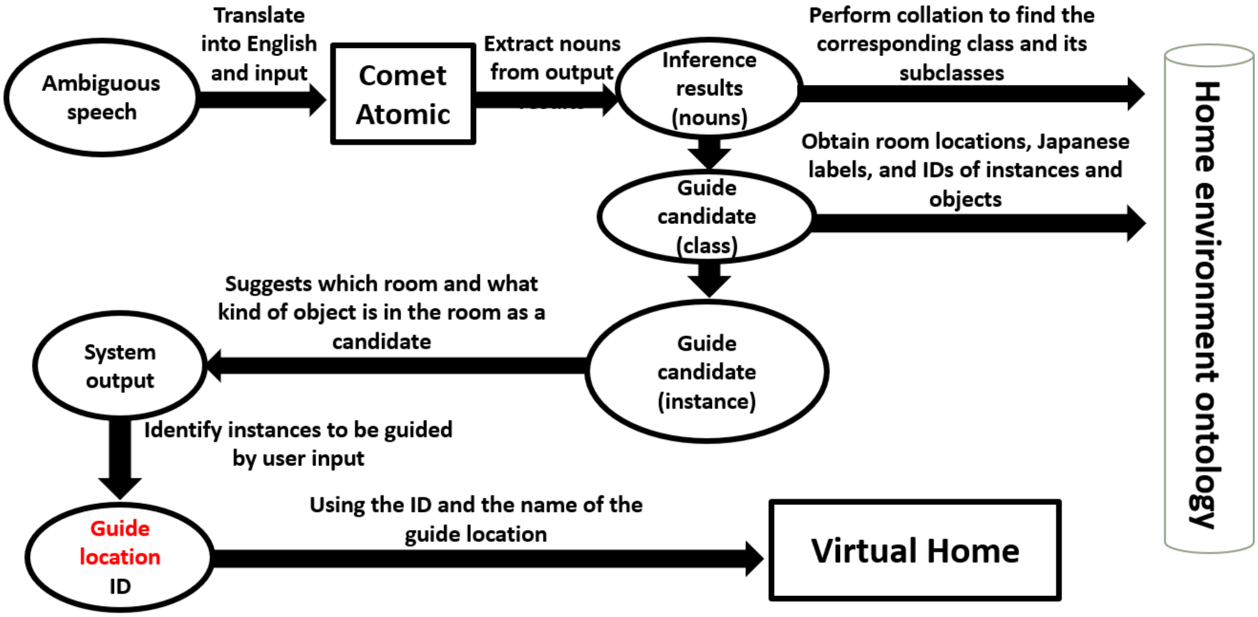
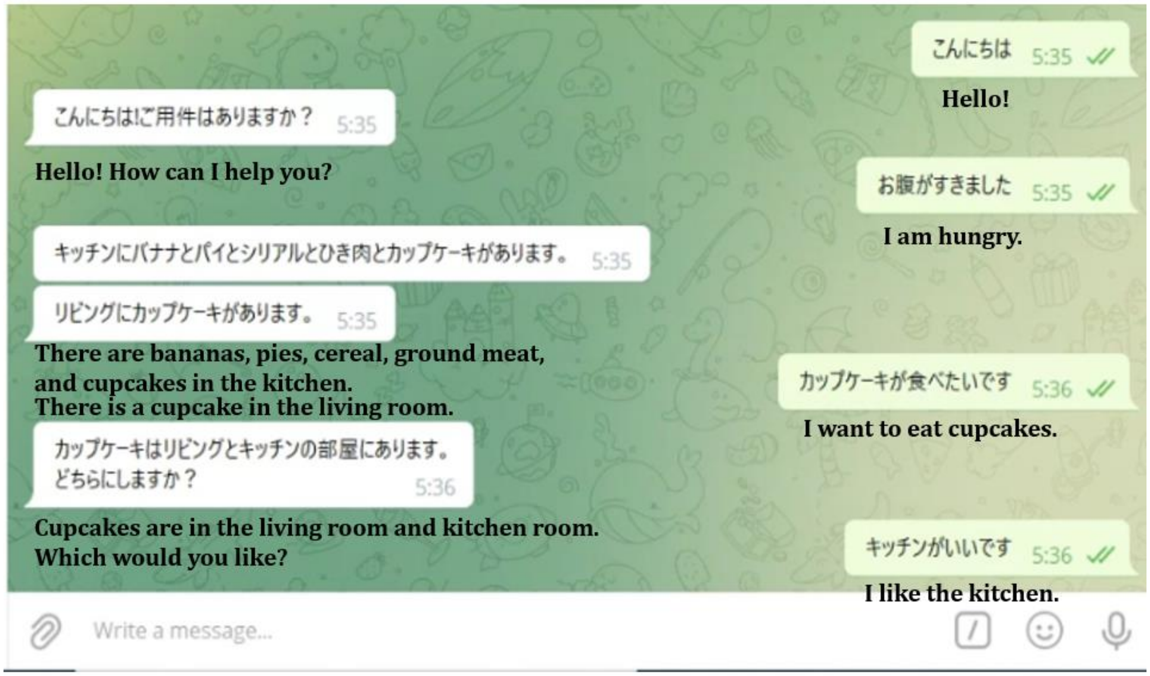

Ambiguous speech
Translate into English and input
Comet Atomic
Extract nouns from output
Inference results (nouns)
Perform collation to find the corresponding class and its subclasses
Home environment ontology
Obtain room locations, Japanese labels, and IDs of instances and objects
Guide candidate (class)
Guide candidate (instance)
Suggests which room and what kind of object is in the room as a candidate
System output
Identify instances to be guided by user input
Guide location ID
Using the ID and the name of the guide location
Virtual Home

Figure 6: The flow of location guidance from user's speech based on



こんにちは 5:35

Hello!

こんにちは!ご用件はありますか？ 5:35

Hello! How can I help you?

お腹がすきました 5:35

I am hungry.

キッチンにバナナとパイとシリアルとひき肉とカップケーキがあります。 5:35

リビングにカップケーキがあります。 5:35

There are bananas, pies, cereal, ground meat, and cupcakes in the kitchen.
There is a cupcake in the living room.

カップケーキが食べたいです 5:36

I want to eat cupcakes.

カップケーキはリビングとキッチンの部屋にあります。どちらにしますか？ 5:36

Cupcakes are in the living room and kitchen room. Which would you like?

キッチンがいいです 5:36

I like the kitchen.

Figure 7: Example of dialogue of the proposed system

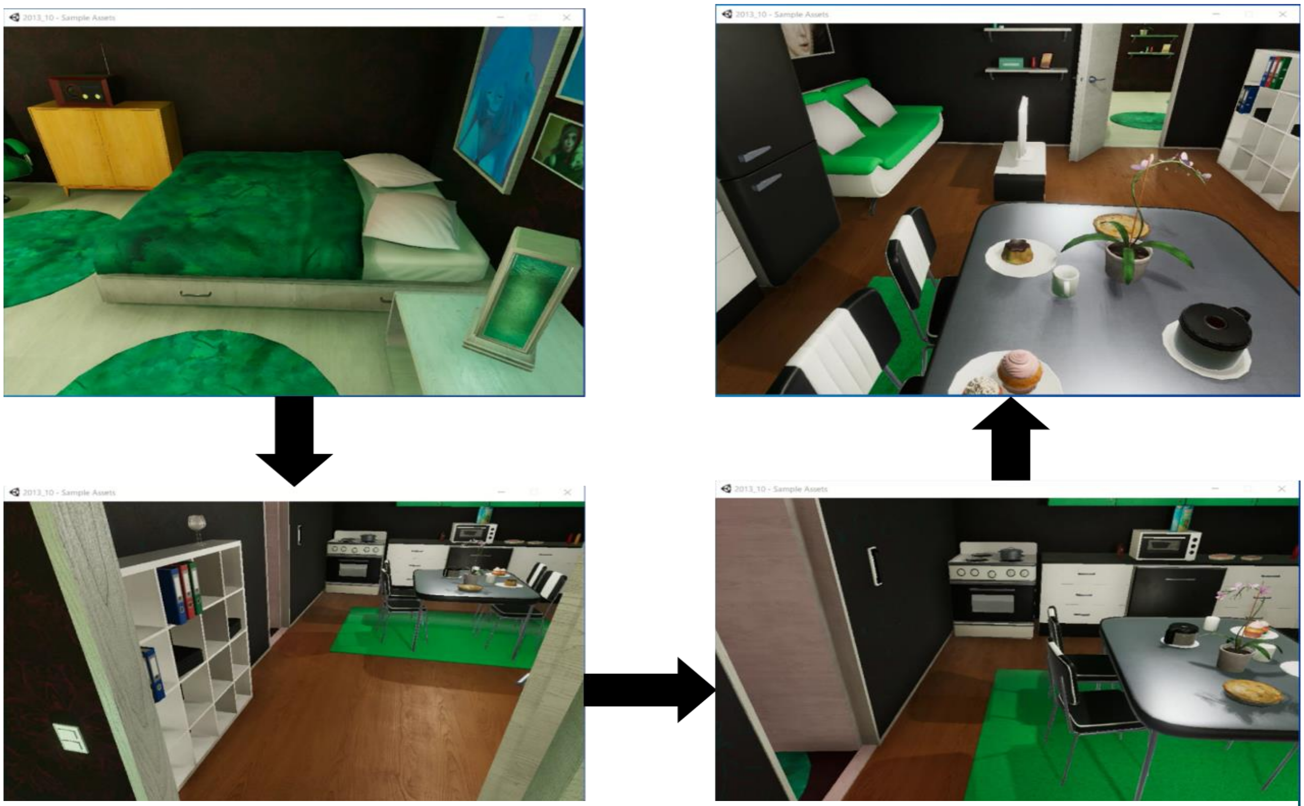


Figure 8: Result of simulation

with a list of candidates, such as "There are cupcakes in the kitchen. The user then interacts with the system to narrow down the location to be guided. When the location is decided, the system obtains the identification number from the instance and guides the user using an action script.

In this way, the system presents several candidates for guidance based on the inference results of the user's ambiguous speech, and the user selects one from them.

4 Execution Example of the Proposed System

Figure 7 shows an example of the proposed system using Telegram. When a user says "I'm hungry" in a dialogue, Rasa's natural language understanding component extracts the user's intention to be hungry from the input sentence. Based on the extracted intention and the dialogue rules, the inference module is used to infer the location of the guide, and the program is executed to identify the location through interaction with the user. The system outputs the information about what is in which room as a candidate for guidance. Next, if the user specifies multiple objects in different rooms, and the system is unable to identify one, it prompts the user to choose which room to be guided to. Finally, if the user is able to identify the room and the location of the object to be guided, the system will guide the user to the desired location through a simulation using VirtualHome, as shown in Figure 8.

5 Policy of the Evaluation Experiment

According to a survey paper on task-oriented dialogue systems [7], the following three methods are used for evaluation experiments.

1. automatic evaluation
   Evaluation methods based on automatic evaluation criteria for components in a system
2. simulation-based evaluation
   Evaluation by building a user simulator using domain knowledge and providing human-like conversations.
3. human evaluation
   Evaluation by having a participant complete a specific task of interacting with the system and evaluating the experience.

In our dialogue system, since the guidance is based on the dialogue with the system on the Embodied AI simulator, human evaluation is necessary to determine whether the guidance is successful or not. The main evaluation items are the task success rate, which indicates whether the guidance is successful or not, the irrelevant turn rate, which indicates whether there are unnecessary dialogues in the process until the success of the task, and the user's satisfaction with the system.

6 Conclusion

In this paper, we described the construction and implementation of an interactive dialog navigation system based on home environment ontology and common-sense reasoning.

The functionality of this system is limited to navigation and does not support other actions provided by VirtualHome. Also, currently, we do not support questions such as How and Why for user input.

In the future, we expect to apply this research to the development of household robots and real estate property guidance robots by reflecting the virtual robots learned in the virtual space in the real world. We plan to add dialogue rules and natural language understanding data sets to the dialogue system to enrich the conversation patterns, and to conduct the evaluation experiments described in Section 5. In addition, after conducting evaluation experiments, we would like to clarify the technical issues and solutions of this system.

We plan to use the Corpus of Everyday Japanese Conversation (CEJC) [10] to expand the dataset of dialogue rules and natural language understanding. the CEJC is a corpus of natural everyday conversations in various situations. In this study, we are considering extending the dataset with a corpus of daily conversations at home in order to conduct dialogues about the home. In addition, although the ontology we have created now provides information about objects that exist in the environment, we will consider using an ontology about activities in the future, as in the research of [12].

REFERENCES

[1] J. Thomason ,et al . Vision-and-Dialog Navigation. CoRL2019.

[2] D. Jena Hwangy , et al. (COMET-)ATOMIC2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI2020.

[3] T. Morita, N. Fukuta,N. Izumi, T. Yamaguchi: DODDLE-OWL: Interactive Domain Ontology Development with Open Source Software in Java, IEICE Transactions on Information and Systems, Special Section on Knowledge-Based Software Engineering, Vol. E91-D No. 4 pp. 945-958 DOI: 10. 1093/ietisy/e91-d. 4. 945. 2018.

[4] C. Fellbaum. WordNet and wordnets. In: K. Brown et al. Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670. 2005.

[5] T. Bocklisch , et al, A. Nichol. Rasa: Open Source Language Understanding and Dialogue Management. NIPS 2017 Conversational AI workshop. 2017.

[6] X. Puig, et al .“VirtualHome: Simulating household activities via pro-grams,” Proc. the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8494–8502. 2018

[7] Z. Zhang, et al. Recent Advances and Challenges in Task-oriented Dialog Systems. Sci. China Technol. Sci. 63, 2011–2027 (2020).

[8] J. Duan, et al. A SURVEY OF EMBODIED AI: FROM SIMULATORS TO RESEARCH TASKS. CVIU2021

[9] M. Burtsev, et al. DeepPavlov: Open-Source Library for Dialogue Systems. Proc. the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pp. 122–127, 2018.

[10] H. Koiso, Y. Iseki, Y. Usuda, W. Kashiwano, Y. Kawabata, Y. Tanaka, Y. Den and K. Nishikawa. Construction of "Japanese Corpus of Daily Conversation". In Proceedings of the 23rd Annual Conference of the Association for Natural Language Processing (NLP2017). (in Japanese)

[11] T. Pejsa, J. Kantor, H. Benko, E. Ofek and A. Wilson: Room2room: Enabling life-size telepresence in a projected augmented reality environment. Proc. of 19th ACM-CSCW, pp. 1716-1725, 2016.

[12] S. Egami, S. Nishimura. and K. Fukuda.: VirtualHome2KG: Constructing and Augmenting Knowledge Graphs of Daily Activities Using Virtual Space. In: International Semantic Web Conference (ISWC) 2021: Posters, Demos, and Industry Tracks. Vol.2980, CEUR-WS (2021)

[13] A. Vassiliades., N. Bassiliades. F. Gouidis and T. Patkos.: A knowledge retrieval framework for household objects and actions with external knowledge. In: International Coanference on Semantic Systems. pp. 36–52. Springer, Cham (2020)