information

MDPI

Article

An Interactive Virtual Home Navigation System Based on Home Ontology and Commonsense Reasoning

Alan Schalkwijk [1], Motoki Yatsu [2] and Takeshi Morita [2,*]

[1] Graduate School of Science and Engineering, Aoyama Gakuin University, Sagamihara 252-5258, Japan; alan.s@kfxbiglobe.ne.jp

[2] College of Science and Engineering, Aoyama Gakuin University, Sagamihara 252-5258, Japan; yatsu@it.aoyama.ac.jp

* Correspondence: 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 research, we aim to develop a two-way interactive navigation system by introducing knowledge-based reasoning to Embodied AI research. Specifically, the system obtains guidance candidates that are difficult to identify with existing common-sense reasoning alone by reasoning with the constructed home ontology. Then, we develop a two-way interactive navigation system in which the virtual robot can guide the user to the location in the virtual home environment that the user needs while repeating multiple conversations with the user. We evaluated whether the proposed system was able to present appropriate guidance locations as candidates based on users' speech input about their home environment. For the evaluation, we extracted the speech data from the corpus of daily conversation, the speech data created by the subject, and the correct answer data for each data and calculated the precision, recall, and F-value. As a result, the F-value was 0.47 for the evaluation data extracted from the daily conversation corpus, and the F-value was 0.49 for the evaluation data created by the subject.

Keywords: navigation system; home ontology; dialog system; common-sense reasoning

check for updates

Citation: Schalkwijk, A.; Yatsu, M.; Morita, T. An Interactive Virtual Home Navigation System Based on Home Ontology and Commonsense Reasoning. *Information* **2022**, *13*, 287. https://doi.org/10.3390/info13060287

Academic Editor: Ryutaro Ichise

Received: 25 February 2022
Accepted: 2 June 2022
Published: 6 June 2022

1. Introduction

In recent years, research has been conducted on Embodied AI [1], which attempts to solve various tasks rooted in complex environments where various objects exist using agents such as virtual robots that learn in the real world or virtual space. In order to solve these tasks in complex environments, it is necessary to integrate technologies from various fields, such as computer vision, computer graphics, natural language processing, artificial intelligence, robotics, and navigation, which are necessary to identify rooms and objects in the environment.

Traditional research on computer vision and natural language processing has been developed by Internet AI, which focuses on pattern recognition of images and text. In contrast, Embodied AI focuses on the ability of a virtual robot to see, hear, speak, and move in a virtual space. Examples of major tasks in Embodied AI include object navigation [2–5], Embodied Question Answering (EQA) [6–9], Vision-and-Language Navigation (VLN) [10], and Vision-dialog Navigation (VDN) [11,12].

Object navigation is a task in which a virtual robot is placed at a random position to find a specific object in the environment. The virtual robot is equipped with two sensors, an RGB-D camera and a GPS compass, and uses the visual information obtained from these sensors and information about the current position and orientation of the virtual robot.

EQA is a task in which a virtual robot randomly placed in a virtual environment moves and perceives the surrounding environment in order to answer the user's questions. The questions asked by the user are related to the rooms and objects in the environment, and the robot can answer questions such as the color of the objects, which room they are in, and what kind of objects are in the room.

VLN is a task where agents learn to navigate the environment by following natural language instructions. VDN is a task in which a human gives unspecified instructions, such as those given in a home environment, and a virtual 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 language tele-operated home robot or an office robot, when the robot does not know where to go next.

In recent years, at the Embodied AI workshop held in conjunction with the CVPR (Conference on Computer Vision and Pattern Recognition) (https://embodied-ai.org/ accessed on 1 June 2022), competitions on the above-mentioned Embodied AI have been held, and various methods and systems are being researched and developed.

In conventional Embodied AI research, tasks such as object navigation and EQA can be solved by a one-way dialogue in which the robot responds to the user's request, such as navigating to the specified object to complete the task. In EQA, the robot responds to the user's question while moving through the environment. In addition, VDN is a task that takes appropriate actions in response to human instructions. However, many of the instructions include specific directions such as "go to bedroom". In addition, it is difficult to identify the location of the guidance system in the environment from ambiguous requests such as "I'm hungry" using only conventional common-sense reasoning. Therefore, in order for the virtual robot to be able to respond to ambiguous requests from the user, it is necessary for the user and the virtual robot to have multiple conversations and to infer an appropriate guidance location using common-sense knowledge and home ontology that incorporates information about the virtual environment.

The objective of this research is to develop a two-sided interactive navigation system by introducing knowledge reasoning techniques to the study of Embodied AI. Specifically, based on knowledge of the home environment 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.

Our contributions are as follows:

1. We proposed a method for presenting candidate guides from ambiguous requests based on home ontology and common-sense reasoning.
2. We proposed a method for constructing home ontology from a VirtualHome environment graph semi-automatically.
3. To evaluate the proposed method, we created a dataset consisting of pairs of "utterances about ambiguous requests in the home" and "objects or rooms in VirtualHome associated with those requests."

The rest of this paper is organized as follows. Section 2 describes the related works, and Section 3 describes the proposed system and its components. Section 4 describes the evaluation data, the evaluation method, and the results and discussion of the evaluation experiments. Finally, the conclusion is given in Section 5.

2. Related Works

In this section, we introduce object navigation, EQA, VDN, a dialogue system related to this research, and research on inference (Commonsense Knowledge Graph, Home Ontology) necessary for the inference of candidate guides.

2.1. Object Navigation

Ref. [2] proposes a modular system called "Goal-Oriented Semantic Exploration", which builds an episodic semantic map and uses it to explore the environment efficiently based on the goal object category. The proposed method achieves state-of-the-art performance on the object goal navigation task and won the CVPR2020 Habitat ObjectNav challenge.

Ref. [3] proposes a simple neural-symbolic approach for object navigation in the AI2-THOR environment [13]. The proposed method takes raw RGB images as input and uses a spatial memory graph as memory to store object and location information. The results demonstrate that the method can perform near-perfect object navigation tasks in a simple kitchen environment.

ObjectGoal Navigation (OBJECTNAV) is an embodied task wherein agents are to navigate to an object instance in an unseen environment. Prior works have shown that end-to-end OBJECTNAV agents that use vanilla visual and recurrent modules, e.g., a CNN+RNN, perform poorly due to overfitting and sample inefficiency. Ref. [4] instead re-enables a generic learned agent by adding auxiliary learning tasks and an exploration reward. The agents achieve 24.5% success and 8.1% Success weighted by Path Length (SPL), a 37% and 8% relative improvement over prior state-of-the-art, respectively, on the Habitat ObjectNav Challenge. Ref. [5] proposes a large-scale study of imitating human demonstrations on tasks that require a virtual robot to search for objects in new environments—(1) ObjectGoal Navigation and (2) PickPlace. Towards this, [5] collects a large-scale dataset of 70k human demonstrations for ObjectNav and 12k human demonstrations for PickPlace tasks using the web infrastructure Habitat-Web (https://github.com/Ram81/habitat-web accessed on 1 June 2022). Ref. [5] uses these data to answer the question: how does large-scale imitation learning (IL) compare to large-scale reinforcement learning (RL)? Overall, this work provides compelling evidence for investing in large-scale imitation learning.

From the above related studies, object navigation studies focus on navigating to an object specified by its label in an unexplored environment [14]. On the other hand, we focus on identifying relevant objects and rooms in the virtual home environment from the user's ambiguous requests through dialogue.

2.2. Embodied Question Answering

Ref. [6] proposes a new AI task: Embodied Question Answering (EQA), where an agent spawned in an environment must intelligently navigate from an egocentric view to gather the necessary information to answer visual questions about its environment. Ref. [6] introduces the EQA dataset of visual questions and answers grounded in House3D [15]. The different question types test a range of agent abilities—scene recognition, spatial reasoning, and color recognition.

Ref. [7] introduces Interactive Question Answering (IQA), the task of answering questions that require an autonomous agent to interact with a dynamic visual environment. There are several key challenges. The agent must be able to interact with objects in the environment (such as opening the microwave, picking up books, etc.). The agent also must be able to plan and execute a series of actions in the environment conditioned on the questions asked of it. Unlike [7], our system is not able to interact with objects in the environment and plan and execute a series of actions in the environment.

Ref. [8] proposes Multi-Target EQA (MT-EQA), extending the original EQA questions from a limited single-target setting to a more challenging multi-target setting, which requires the agent to perform comparative reasoning before answering questions. In MT-EQA, Ref. [8] proposes six types of compositional questions that compare attribute properties (color, size, distance) between multiple targets (objects/rooms). Unlike [8], our system does not consider questions that have multiple targets.

Ref. [9] investigates a new AI task—Multi-Agent Interactive Question Answering—where several agents explore the scene jointly in interactive environments to answer a question. To cooperate efficiently and answer accurately, agents must be well-organized to have balanced work division and share knowledge about the objects involved. Unlike [9], this study does not use multiple agents.

The above studies cover more complex questions including multiple objects than this study, but those questions include specific labels for objects and attributes. On the other hand, our system differs from the related studies in that it can respond to ambiguous requests that do not directly include labels for objects or rooms in the virtual home environment.

2.3. Vision-Dialog Navigation

Robots navigating in human environments should use language to ask for assistance and be able to understand human responses. To study this challenge, Ref. [11] proposes Cooperative Vision-and-Dialog Navigation (CVDN), an English language dataset situated in the Matterport Room-2-Room (R2R) simulation environment [16].

Vision-dialog navigation (VDN) also requires handling the language intentions of a series of questions about the temporal context from dialogue history well and co-reasoning both dialogs and visual scenes. Ref. [12] proposes a Cross-modal Memory Network (CMN) to exploit the agent memory about both the linguistic interaction with the human and the visual perception from the environment in the task of Dialog History (NDH) where the dialogs between the navigator and the oracle are pre-annotated.

The main dialogue examples in VDNs 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. The goal of this study is to estimate and guide the user to the location based on more ambiguous and everyday speech such as "I'm hungry", as shown in Figure 1, rather than specific instructions such as "Go to the room with the bed", which is assumed in VDNs.

In addition, in the related studies, the user speaks to the system, and the virtual robot responds and takes actions appropriate for the task. In this study, the challenge is to achieve a two-way dialogue between the user and the virtual robot and to clarify the location of the guidance by having the virtual robot also talk about the virtual home environment, as shown in Figure 1. As an approach to this issue, this study attempts to estimate the guide location using common-sense reasoning and home ontology.

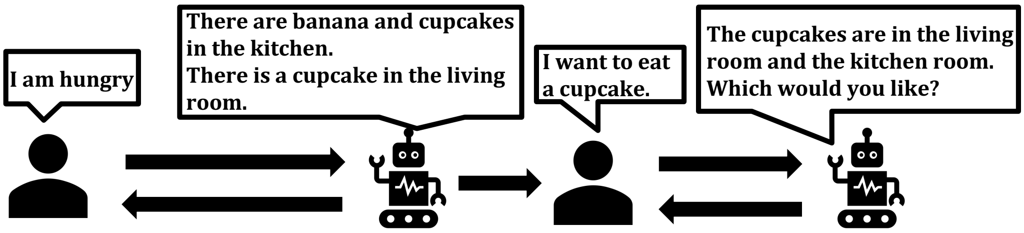

I am hungry
There are banana and cupcakes in the kitchen.
There is a cupcake in the living room.
I want to eat a cupcake.
The cupcakes are in the living room and the kitchen room.
Which would you like?

Figure 1. Example of interaction with the system.

2.4. Common-Sense Knowledge Graph

Previous research on common-sense knowledge include COMET-ATOMIC [17], a common-sense 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".

2.5. Home Ontology

A knowledge retrieval framework for household objects and actions with external knowledge [18] was created known as a home ontology built on VirtualHome. VirtualHome contains a dataset of labels for human household actions and their textual and action script representations. VirtualHome provides a dataset that consists of a set of labels, sentences, and action scripts that describe human behavior in the home and objects that exist in the home. Using this dataset, we have created an ontology of a series of actions and the objects that appear in those actions. We have also constructed SPARQL queries for this ontology, which can be used to answer questions about behaviors and objects in the home.

In this study, we use the COMET-ATOMIC language model introduced in Section 2.4 to extract nouns from its output that can be used to infer the target location for everyday ambiguous speech. In order to identify which objects in the home environment can be guided by class matching, we have created a new home ontology base that focuses on the relationship between objects and rooms in a VirtualHome. In addition, the home ontology created in this research has instances for all the properties and objects that exist in the environment that are necessary for guidance. The details of the home ontology are described in Section 3.5.

Another study of VirtualHome's home ontology is [19]. Ref. [19] provides a system that can simulate the activities of a virtual robot using VirtualHome, record the spatio-temporal changes in the virtual space, and construct or extend an ontology of daily life activities based on the data.

3. Proposed System

In this section, we describe the structure of the proposed system, the construction procedure of the home ontology, and the dialogue based on common-sense reasoning. In the proposed system, a user inputs a sentence into 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). In this section, Section 3.1 describes the system configuration, Section 3.2 describes the interface, Section 3.3 describes the dialogue system, Section 3.4 describes the Embodied AI Simulator, Section 3.5 describes home ontology, Section 3.6 describes knowledge-based reasoning, and Section 3.7 presents an example run of the proposed system.

3.1. System Configuration

Figure 2 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 ontology and common-sense reasoning to infer the location of the guidance from the speech. Next, we discuss the rationale for selecting the frameworks and models used to create each module. First, for the interactive system, we surveyed the latest papers on dialogue system frameworks [20–23] and compared Rasa with other frameworks. As a result, we found that Rasa is superior for the following reasons:

- It can extract intentions and entities from user input sentences;
- Its 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.

Second, we describe our reasons for selecting Embodied AI. We compared Embodied AI simulators [13,24–31] by referring to a survey paper on the latest Embodied AI simulators [1]. We selected our Embodied AI simulators based on the following criteria: the richness of their object states and inter-object properties and their ability to provide a home-like environment. Among the simulators, we decided to use VirtualHome because it represents the home environment using an environment graph, and it is easy to use to create knowledge about the home environment. Finally, knowledge-based reasoning uses COMET-ATOMIC, which provides a model capable of making the common sense inferences needed to infer guide locations from ambiguous utterances.

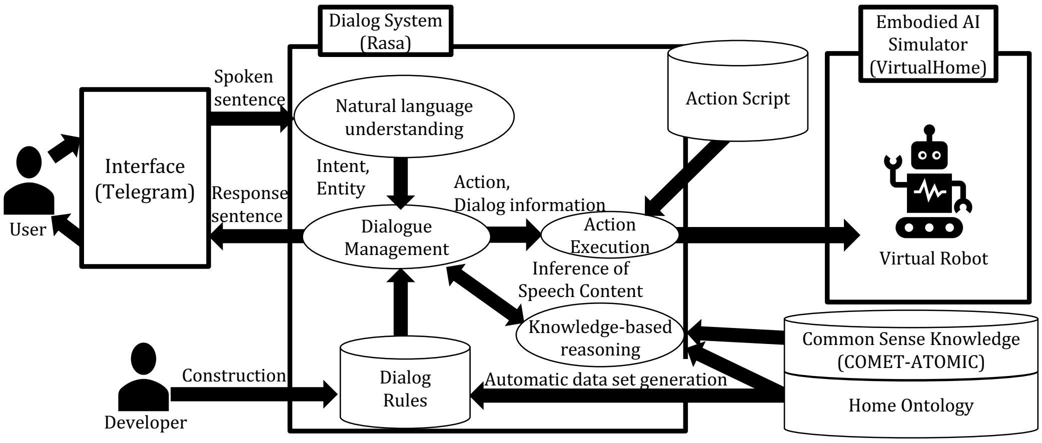


Figure 2. System configuration diagram.

3.2. Interface

This module is positioned at the interface of Figure 2. In this study, we use Telegram (https://telegram.org accessed on 1 June 2022) as an interface because Telegram is an instant messaging application that allows users to have conversations by text input. Telegram is used as a dialogue interface in this study because it is easy to introduce and is used by many people and because it can be linked with the dialogue system created by Rasa [32], the dialogue framework used in this study. Ngrok (https://ngrok.com accessed on 1 June 2022) is used for communication between Telegram and the dialogue system, as shown in Figure 3. Ngrok is a service that can expose a network service running on a local PC to the outside world. The output of the dialogue system is exposed to the outside world using ngrok and is sent to Telegram, the interface, via the net. Similarly, when a user sends input to Telegram, it is sent to the dialogue system via the net, thus enabling collaboration.

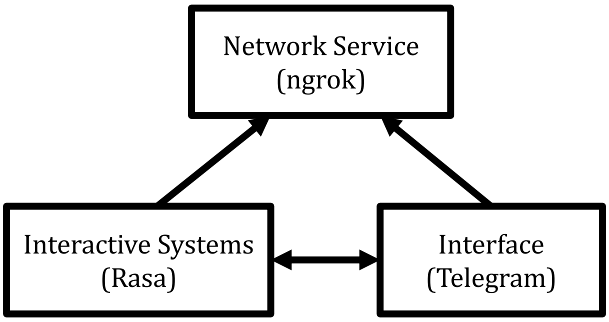


Figure 3. Linking dialogue systems and interfaces.

3.3. Dialogue System

This module is positioned in Figure 2's dialogue system.

3.3.1. Overview

To build this dialogue system, we use Rasa [32], an open-source dialogue framework for developing task-oriented dialogue systems, which is roughly divided into a natural language understanding module and a dialogue management module. The natural language understanding module shown in Figure 4 prepares a dataset of example dialogs corresponding to user-created intentions, and through training, it is able to extract intentions and entities from the user input. Then, the system performs dialogue based on the dialogue rules created by the dialogue management module. The dialogue rules can determine the output of the system in response to the intentions extracted from the user's input.

In the example shown in Figure 4, when the user first says "I'm hungry", the Natural Language Understanding module extracts the intention "ambiguous", which is the closest example of the utterance "I'm hungry". After the intention is extracted, the dialogue management module decides which dialogue rule to execute. Here, there is a rule called "-intent: ambiguous", which is described under "steps:" and ambiguous. In this section, the rule "-intent: ambiguous" is described under "steps:", and if the intention "ambiguous" is extracted, the action of the virtual robot described by "-action:" (hereinafter referred to as "action") in the following line is executed. The actions of the virtual robot described by "-action:" (hereinafter referred to as "action") include the execution of a specific speech or a specific program written in Python by the user.

When executing an original program, as shown in Figure 5, a server called the "Rasa Action Server" can be used. When a user needs to perform an original action according to the dialogue rules, he or she can send to the server the action of the virtual robot described in action, as well as system information such as user input, extracted intentions, and entities. Using this information, a Python program created by the user can be executed and the results returned to the dialogue system. For these reasons, we decided to use Rasa for the development of the dialogue system in this study. In the next section, we introduce the natural language understanding module, the dialogue management module, and the system's speech behavior.

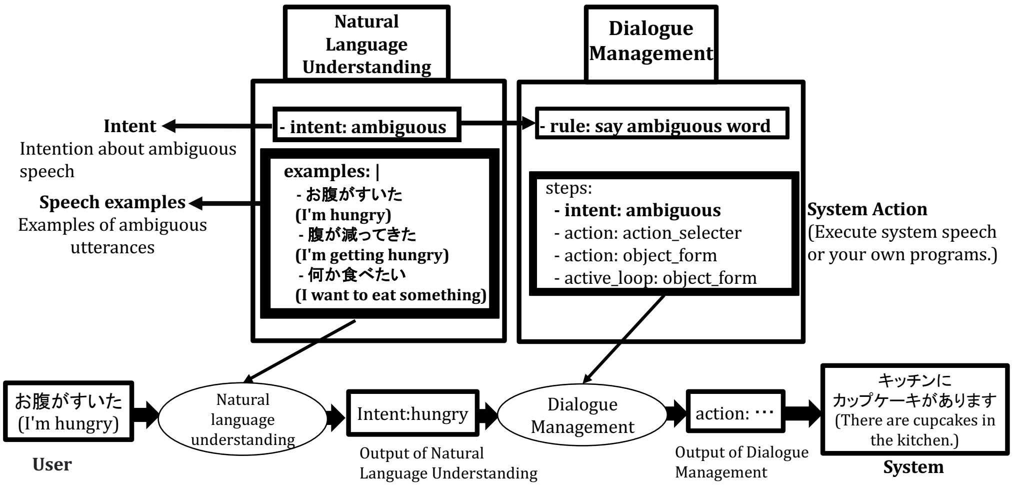

Natural Language Understanding
Dialogue Management
Intent
Intention about ambiguous speech
- intent: ambiguous
- rule: say ambiguous word
Speech examples
Examples of ambiguous utterances
examples: |
- お腹がすいた
(I'm hungry)
- 腹が減ってきた
(I'm getting hungry)
- 何か食べたい
(I want to eat something)
steps:
- intent: ambiguous
- action: action_selecter
- action: object_form
- active_loop: object_form
System Action
(Execute system speech or your own programs.)
お腹がすいた
(I'm 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 4. Example of system input/output.

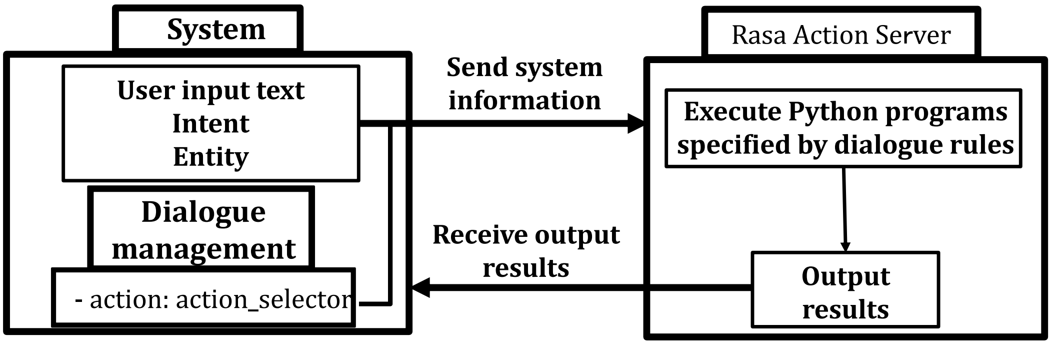
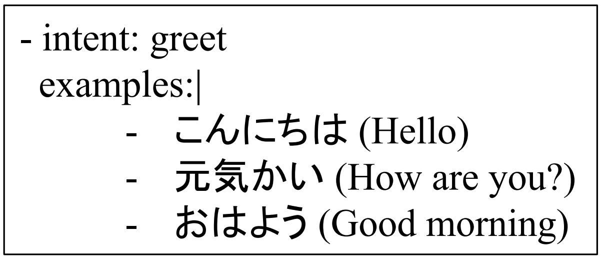


Figure 5. Details of Rasa Action Server.

3.3.2. Natural Language Understanding Module

In this section, we describe the natural language module that we created in this research. A natural language module is a module that extracts speech intentions and entities from user speech. As a dataset for learning with the natural language module, we created four intentions and 200 examples of utterances for these intentions. An example of the dataset is shown in Figure 6. The user′s intentions are described after "-intent:", and examples of utterances in response to the intentions are described below "example: ∣". Table 1 shows some examples of intentions and their corresponding speech. In the example of the "guide" intention, the part described in "[kitchen]"entity": "guidename"" can be extracted as an entity. It is possible to extract the part described in [] as an entity.

For example, when a user says "Please guide me to the bath", "bath" is extracted as the entity name "guidename". The extracted entities can then be used in programs created by the user via the Rasa Action Server.



```
- intent: greet
  examples:|
        -   こんにちは (Hello)
        -   元気かい (How are you?)
        -   おはよう (Good morning)
```

Figure 6. Example dataset description of the natural language module for greeting intentions.

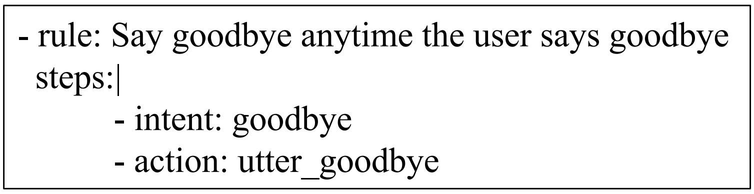
Table 1. List of Natural Language Module Data Set Description.

Intention	Intention Details	Speech Examples	Number of Speech Examples
greet	User greeting utterances	Hello	10
goodbye	User′s utterances about farewell speech	Goodbye	10
guide	Intentions about utterances from which entities can be extracted and from which no inference needs to be made	Please guide me to [kitchen] {"entity": "guidename"}	15
ambiguous	Intentions about ambiguous speech	I'm hungry I'm thirsty I'm a little tired I'm getting sleepy I need somewhere to sit	200

3.3.3. Dialogue Management Module

In this section, we describe the dialogue management module created in this research. The dialogue management module determines the actions of the virtual robot based on the intentions extracted by the natural language module. In the dialogue management module, we created six dialogue rules as data sets. The description of the dialogue rules is shown in Figure 7. After "- rule:" is the name of the rule to be created, and after "steps:", the dialogue rules are written in order. When the intention specified in "- intent:" arrives, the virtual robot will execute the action specified in "- action" below it. The details of the dialogue rules and the actions of the virtual robot are shown below:

1. Rules for greeting rule:
   When the intention "greet" is extracted, the robot's behavior is to return the greeting.
2. Rules for farewell greeting:
   When the intention "goodbye" is extracted, the robot's behavior returns the greeting.
3. Rules for the case when the classification of intentions is not possible:
   When the intention classification score is low in the natural language understanding component, i.e., the intention cannot be extracted from the user's speech, the intention "nlu_fallback" is returned, and the system responds with content that does not understand the user's speech.
4. Rules for greeting the user when the system starts up:
   The system greets the user when the system starts up.
5. Rules for ambiguous speech:
   This rule is used when an ambiguous utterance such as "I am thirsty". is made, and "ambiguous" is extracted by the Natural Language Understanding module. This rule uses the reasoning module to find the best guidance candidate for the ambiguous utterance and suggests it to the user. Then, after multiple conversations with the user, the system identifies the guidance location and guides the user on VirtualHome. This sequence of events is written in Python and executed through the Rasa Action Server.
6. Rules for cases where the name of the location is included and there is no need for inference:
   If the name of the location is extracted as an Entity in the speech, it is used to identify the location by class matching with the home ontology, and if the Entity is not extracted well, the location is inferred using the inference module.



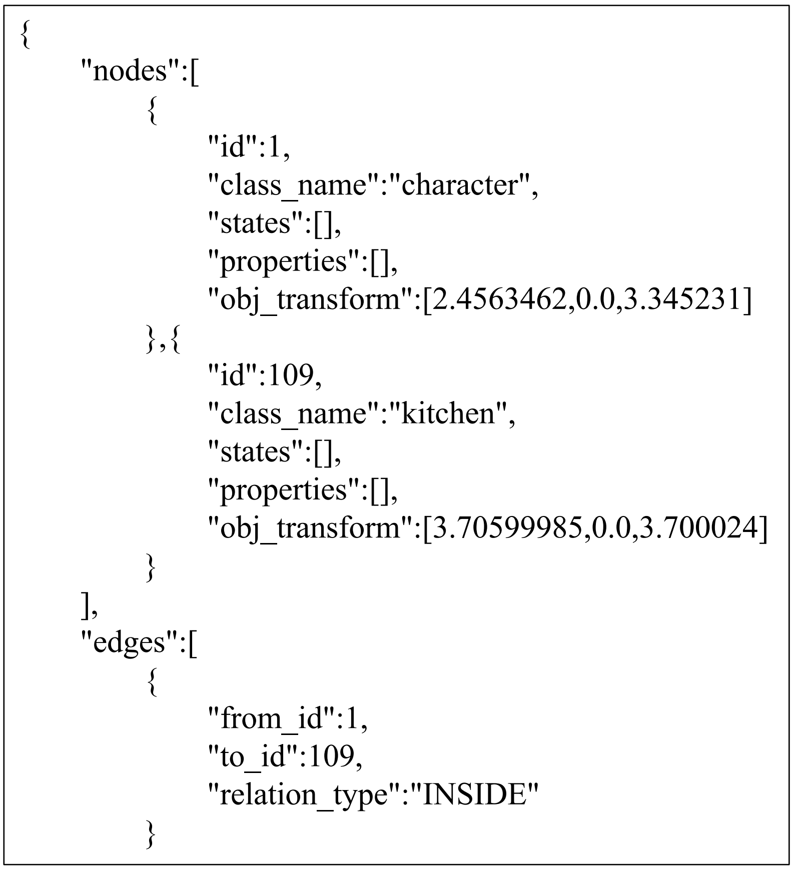
```
- rule: Say goodbye anytime the user says goodbye
  steps:|
        - intent: goodbye
        - action: utter_goodbye
```

Figure 7. Example of dialogue management module dataset description.

3.4. Embodied AI Simulator

This module is positioned in Figure 2's Embodied AI Simulator. The Embodied AI simulator simulates the behavior of a virtual robot in a virtual environment. We use VirtualHome [19] for the Embodied AI simulator, which aims to simulate the activities in a virtual home. The environment of VirtualHome is represented as the environment graph shown in Figure 8. Semantic data such as rooms and objects, their identification numbers, location coordinates, and object states are defined. In addition, the relationship between each object described by "nodes" is represented by "edges", which are defined as "from_id" and "to_id". The identification number of the object is specified by "from_id" and "to_id", and the relationship between the two objects is described by "relation_type". For example, Figure 8 shows that "character" (identification number 1) is inside "kitchen" (identification number 2). In this research, we use this environmental knowledge to construct a home ontology.

In VirtualHome, specific tasks can be performed using action scripts. The action script is shown in Figure 9. It 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. For the action of the virtual robot, only [Walk] is used in this study since it is assumed that the robot moves to guide the user to the target location. The names and identification numbers of the rooms and objects to be acted upon are prepared as variables in the program, and when the guidance location is identified through conversation with the user, the names and identification numbers of the guidance location are referred to from the home ontology and assigned to the respective variables for use. For example, if the user wants to be guided to the living room through a conversation with the system, the system refers to its knowledge of the home environment and obtains the name of the guidance location, "living_room", and the identification number, 1. Then, we assign them to the variables Object_Location and Object ID, respectively, and finally create an action script named "[Walk] <living_room> (1)". Then, we create a script called "[Walk] <living_room> (1)" and make the virtual robot move to the desired location.



```
{
    "nodes":[
        {
            "id":1,
            "class_name":"character",
            "states":[],
            "properties":[],
            "obj_transform":[2.4563462,0.0,3.345231]
        },{
            "id":109,
            "class_name":"kitchen",
            "states":[],
            "properties":[],
            "obj_transform":[3.70599985,0.0,3.700024]
        }
    ],
    "edges":[
        {
            "from_id":1,
            "to_id":109,
            "relation_type":"INSIDE"
        }
```

Figure 8. Part of the environmental graph.

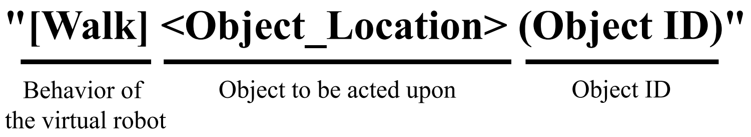

"[Walk] <Object_Location> (Object ID)"
Behavior of the virtual robot
Object to be acted upon
Object ID

Figure 9. Example action script.

3.5. Home Ontology

This module is positioned in Figure 2's Home Ontology. There is prior work on ontologies for behavior in VirtualHome. 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 10 shows the construction procedure of the home ontology. In Step 1 of Figure 10, 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 [33] and WordNet [34]. In Step 3, Japanese labels are added to the added classes in order to perform Japanese dialogues. In Step 4, all environment data are 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. Owlready [35] is used for the ontology operations performed in these steps.

In the end, the number of classes, properties, and instances of the home ontology created was 239, 6, and 356, respectively. Figure 11 shows part of the class hierarchy of the created ontology, and Table 2 shows the properties and their domains, ranges, and descriptions. The namespace of each property shown in Table 2 uses the properties that exist in schema.org. Furthermore, an example of the created instance is shown in Figure 12. This ontology is written in RDF/XML format.

Table 2. Properties and their domains and ranges.

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

It is possible to identify candidate guides by matching the results of common sense inference with the class hierarchy in Figure 11 and obtaining instances of the matched classes. By using the property "containedInPlace" in Table 2, which indicates in which room the identified candidate guide is located, it is possible to show the user in which room the candidate guide is located. It is also possible to guide the user by obtaining the object ID and its coordinates from the property from the candidate guide determined by the interaction with the user.

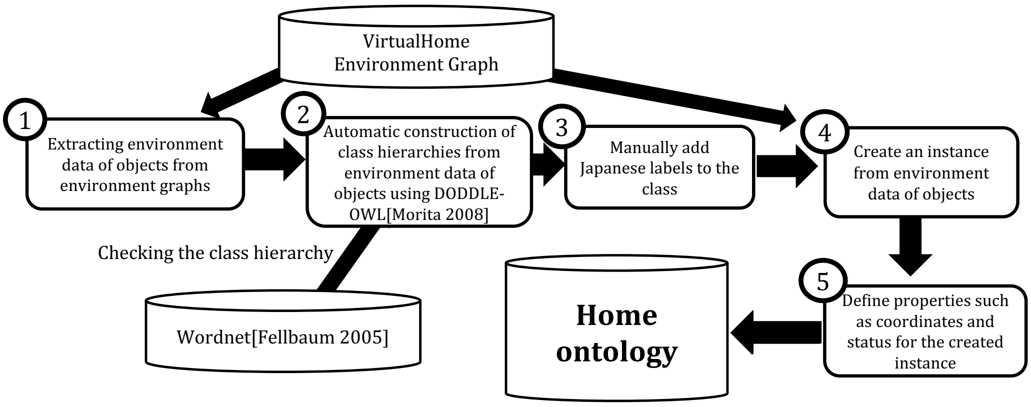

VirtualHome Environment Graph
1
Extracting environment data of objects from environment graphs
2
Automatic construction of class hierarchies from environment data of objects using DODDLE-OWL[Morita 2008]
3
Manually add Japanese labels to the class
4
Create an instance from environment data of objects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

Figure 10. Home ontology construction procedure.

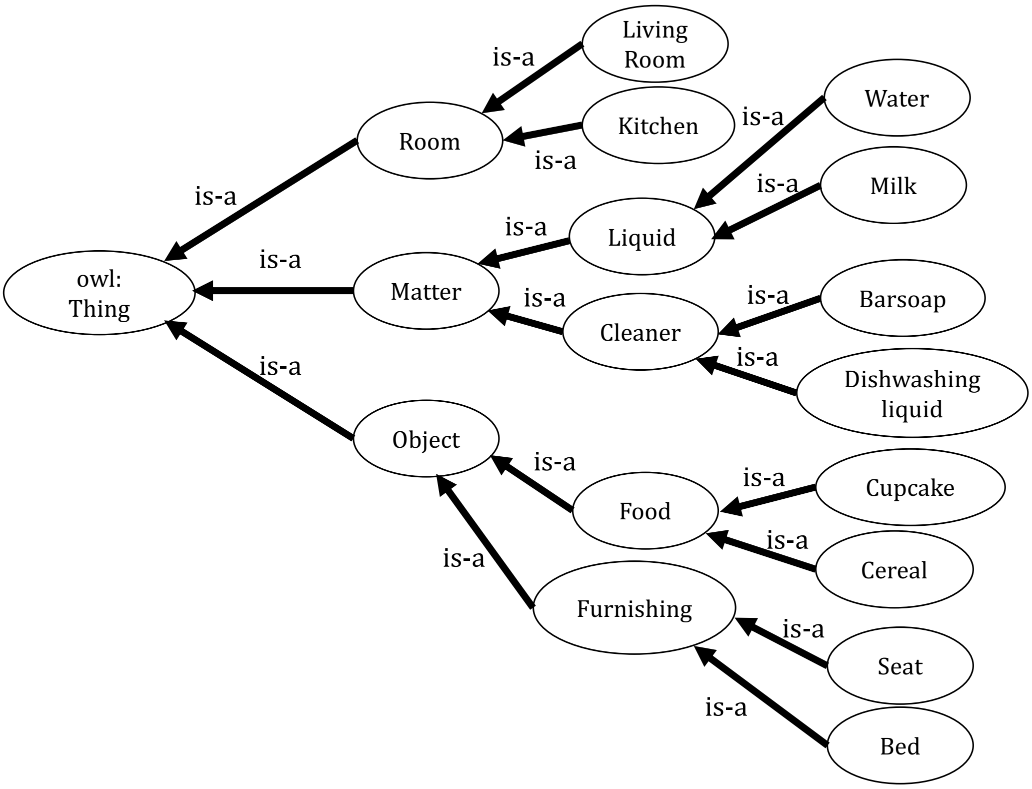
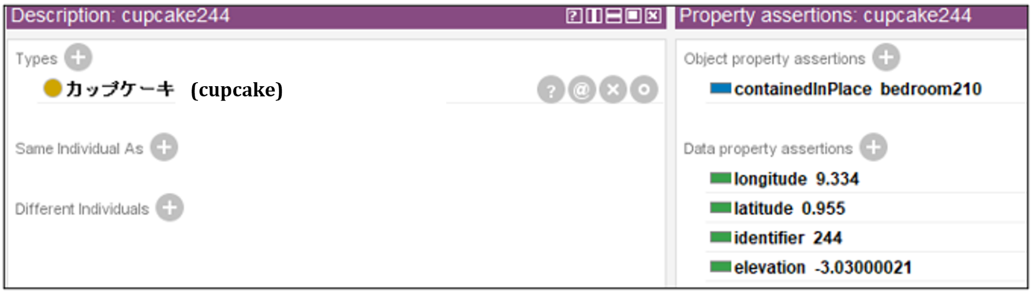


Figure 11. Part of class hierarchy.


Description: cupcake244

Types

● カップケーキ (cupcake)

Same Individual As

Different Individuals

Property assertions: cupcake244

Object property assertions

containedInPlace bedroom210

Data property assertions

longitude 9.334

latitude 0.955

identifier 244

elevation -3.03000021

Figure 12. Examples of instances in home ontology.

3.6. Knowledge-Based Reasoning

This module is positioned in Figure 2's knowledge-based reasoning. In the proposed system, the knowledge-based reasoning is used to reason about ambiguous utterances and identify appropriate places to go. We use the home ontology and COMET-ATOMIC for common-sense reasoning. The flow from user speech to location guidance based on the home ontology and common-sense reasoning is described in Figure 13 and examples.

When a user inputs an ambiguous sentence such as "I am hungry" through Telegram, COMET-ATOMIC, which was introduced in a related study [17], translates the sentence into English as "I am hungry" in order to perform inference. Next, the translated English sentence is input to COMET-ATOMIC, and the system uses three common-sense relationships.These common-sense relationships use 23 types of common-sense relationships, check the results of the acquired guidance candidates against the COMET-ATOMIC input/output and home ontology, and select the ones with appropriate results.

1. xNeed
   This relationship means "actions that humans need to take in response to certain events";
2. xEffect
   This relationship means "events that humans cause in response to certain events";

3. AtLocation
   This relationship means "places where certain objects may be found".

It 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 ontology. From the instances of the matched classes, the system obtains the property values and Japanese labels related to the location of the room and presents them to the user as candidates, such as "There are cupcakes in the kitchen". The user then communicates with the system via Telegram to narrow down the location to be guided. If the user decides on a location, the system obtains an identification number from the instance and guides the user using an action script.

In this way, the system presents several guidance candidates based on the inference results of the user's ambiguous utterances, and the user selects one from them, making it possible to provide guidance using interactive dialogue.

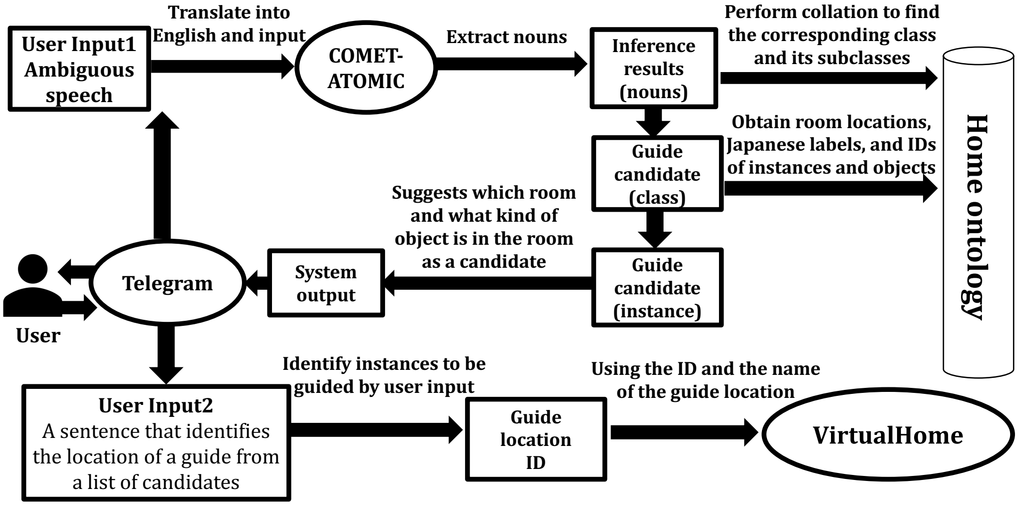
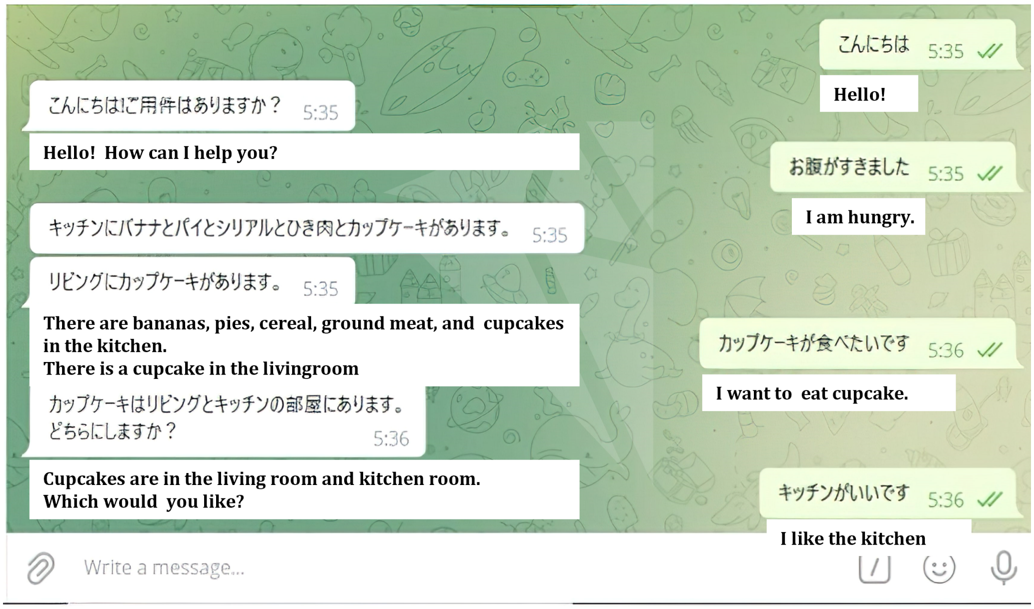

User Input1 Ambiguous speech
Translate into English and input
COMET-ATOMIC
Extract nouns
Inference results (nouns)
Perform collation to find the corresponding class and its subclasses
Home ontology
Guide candidate (class)
Obtain room locations, Japanese labels, and IDs of instances and objects
Guide candidate (instance)
Suggests which room and what kind of object is in the room as a candidate
System output
Telegram
User
User Input2
A sentence that identifies the location of a guide from a list of candidates
Identify instances to be guided by user input
Guide location ID
Using the ID and the name of the guide location
VirtualHome

Figure 13. The flow of location guidance from user's speech based on.

3.7. Execution Example of the Proposed System

Figure 14 shows an example of dialogue in Japanese with the proposed system using telegrams and its English translation. When the input "I'm hungry" is given 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. When the room and the location of the object to be guided are finally identified, the system will guide the user to the desired location through a simulation using VirtualHome, as shown in Figure 15.



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 livingroom

カップケーキが食べたいです 5:36

I want to eat cupcake.

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 14. Example of dialogue of the proposed system.

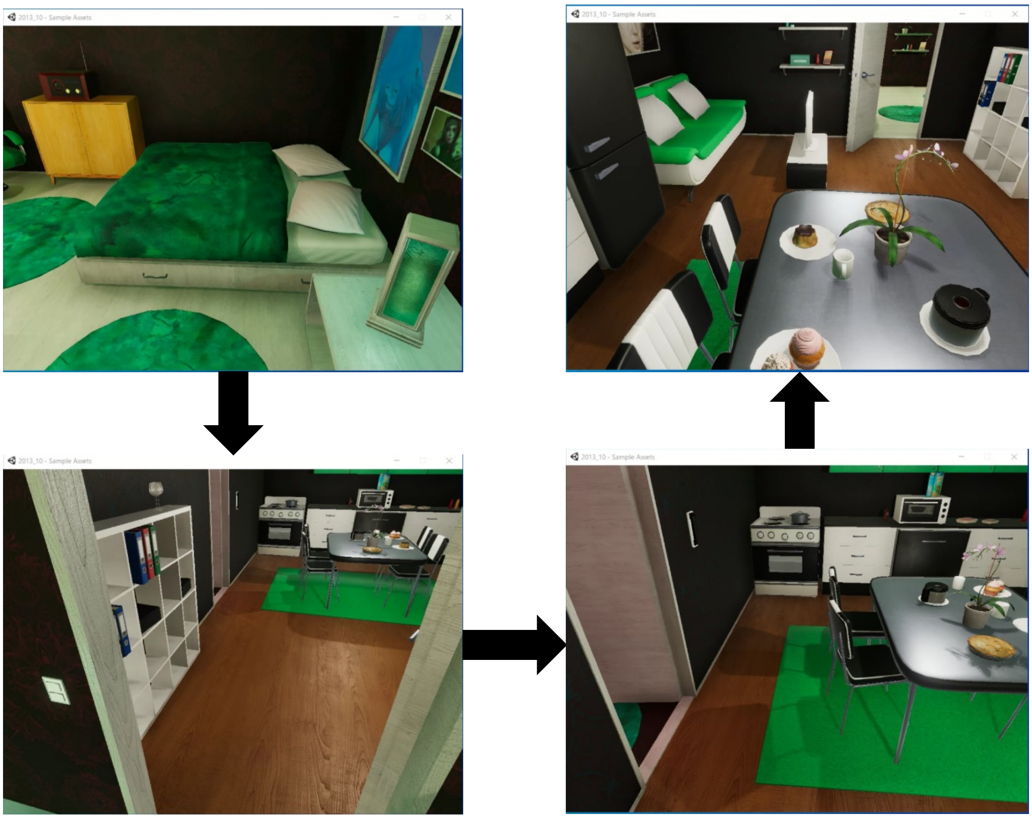


Figure 15. Result of simulation.

4. Evaluation

4.1. Outline of the Evaluation Experiment

In this evaluation experiment, we evaluate whether the system is able to present an appropriate guidance location as a candidate based on the user's speech input. In order to evaluate the proposed system, we need input data that includes everyday speech in a home environment. Since no existing dataset contains utterance data for the home environment, we create a new dataset for the evaluation of the proposed system. Section 4.2 describes the method for creating the evaluation data, Section 4.3 describes the evaluation method, and Section 4.4 describes the results and discussion of the evaluation experiment.

4.2. Data for Evaluation

Using the Corpus of Japanese Everyday Conversation (CEJC) [36], which collects speech data on daily conversation, we generated 100 speech examples for evaluation as input to our dialogue system using the procedure shown in Figure 16.

The first step is to obtain the classes related to objects in the home ontology and their superclasses. Next, we extract all the Japanese labels from the classes we obtained. In addition, we manually create human demands, states, and behaviors that are predicted from each class. The extracted Japanese labels and words related to demands and states are used as search candidates, and string searches are performed on the corpus of Japanese daily conversation. From the search results, we select conversations that can be used in the evaluation dataset. In addition, punctuation marks, personal names, metadata, etc., included in the selected conversations are removed, and the data are processed as a final evaluation dataset.

In addition, we asked three subjects to cooperate in creating 100 utterances related to daily conversation by referring to Figure 17 of the home environment and a list of rooms and objects (shown in Table 3) that exist in the home environment and by predicting possible utterances in that home environment. In the end, a total of 200 speech data extracted from the corpus of Japanese daily conversation, and the speech data created by the subjects were used as the data for evaluation.

Then, we asked the subjects to cooperate and select from among the choices of rooms and things in the virtual home environment of VirtualHome that were related to each utterance in the evaluation data. The selected data for all the subjects were combined to form the correct answer data.

The dataset created and some of the results of the experiment are shown in Table A1 in Appendix A.

Table 3. Part of the objects and room options present in the home environment used in the evaluation experiment.

Types Present in the Home Environment	Name
Room	bathroom, bedroom, kitchen, livingroom
Object	cupcake, plate, toilet paper, towels, keyboard, washing machine, mouse, sewing kit, etc.

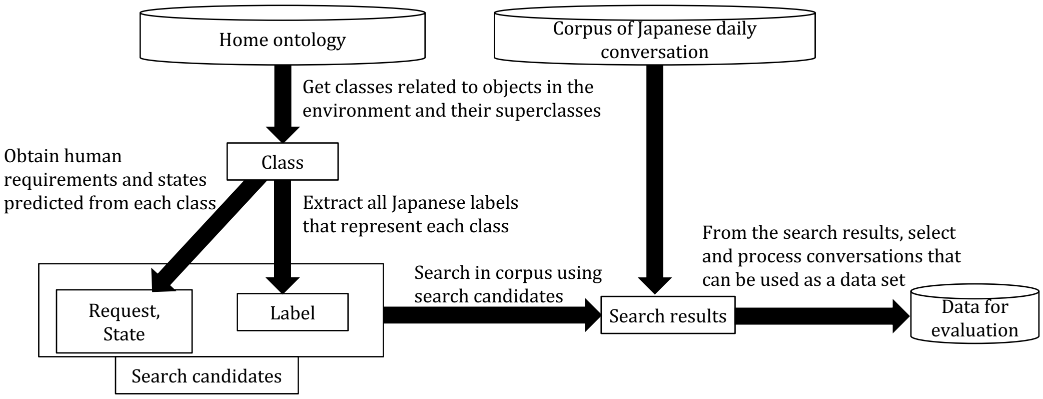

Home ontology
Corpus of Japanese daily conversation
Get classes related to objects in the environment and their superclasses
Class
Obtain human requirements and states predicted from each class
Extract all Japanese labels that represent each class
Request, State
Label
Search candidates
Search in corpus using search candidates
Search results
From the search results, select and process conversations that can be used as a data set
Data for evaluation

Figure 16. Procedure for creating a data set for evaluation.

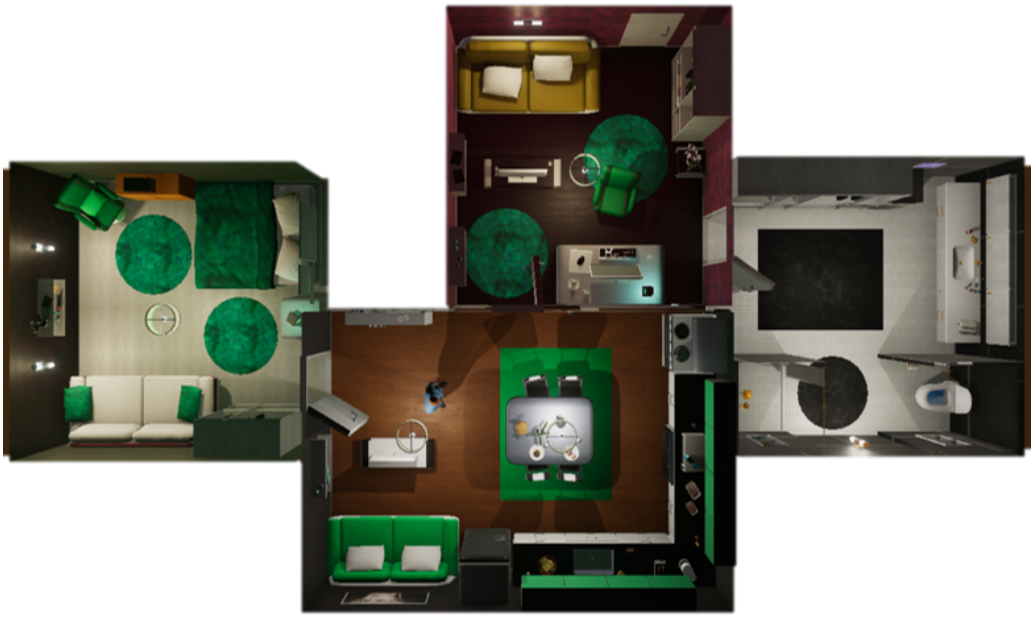


Figure 17. Home environment used in the evaluation experiment.

4.3. Evaluation Method

We input the utterances of the evaluation dataset into the proposed system, and using the output results and the correct answer data, we obtain the precision, recall, and the F-value based on Equations (1)–(3), which are used as the evaluation values.

In Equation (1), the number of candidates that are correct answers in the guidance candidates output by the system is divided by the number of all guidance candidates to obtain the precision. In Equation (2), the number of candidates that are correct answers in the guidance candidates output by the system is divided by the number of all correct answers generated by the subject, and the recall is obtained. In Equation (3), the F-value is calculated using the precision and recall rate calculated by Equations (1) and (2).

$$\text{Precision} = \frac{\text{The number of correct answers in the system output}}{\text{The number of system outputs}} \tag{1}$$

$$\text{Recall} = \frac{\text{The number of correct answers in the system output}}{\text{The number of all correct answers}} \tag{2}$$

$$\text{F-value} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

4.4. Experimental Results

Table 4 shows the averages of the precision, recall, and F-value of the evaluation data for each speech case. In this evaluation experiment, the F-values of the corpus of everyday Japanese conversation and the examples of speech produced by the test subjects were 0.47 and 0.49, indicating that there was not much difference between them. The reason for this is that the corpus of Japanese daily conversation and the speech examples created by the test subjects are both related to daily life in the home environment, and there is not much difference in the data sets. In addition, the F-values of the corpus are 0.47 and 0.49 in both cases, suggesting that the system was able to provide users with some appropriate guidance candidates.

Possible reasons for not being able to answer the user’s requests include the following: the COMET-ATOMIC output results did not contain nouns and could not be successfully matched with the home ontology due to the lack of noun extraction. The extracted nouns were not present in the class of the home ontology; the sentences entered into COMET-ATOMIC were either too long or too short or were lacking a subject or object, and the results could not be qualified.

Table 4. Average of precision, recall and F-value.

Evaluation Data Set	Precision	Recall	F-Value
Japanese Corpus of Everyday Conversation	0.48	0.58	0.47
Examples of utterances created by subjects	0.48	0.62	0.49

4.5. Discussion

In this section, we will discuss the causes of the failure of the system to properly output the guidance candidates based on the correct answers and failures in the actual evaluation experiment results.

Table 5 shows the examples of speech translated into English, the system output, and the correct answer data. Table 6 shows the input of COMET-ATOMIC, the English translation, the output results, the results of class matching with the home ontology by extracting nouns from the output results, and the Japanese labels of the matched classes. From Table 6, we can see that the two candidate Japanese class labels for the system's output are "bedroom" and "bed", which are used to obtain the "system output" shown in Table 5. The precision and the recall are calculated using the class labels that are candidates for guidance in the underlined system output and the correct answer data. The precision is calculated using Equation (1) in the previous section. In the case of Table 5, the number of correct answers in the system output is "bedroom" and "bed", and the number of system output will be two, "bedroom" and "bed". The above two numbers were substituted into Equation (1), and the result had a goodness of precision of 1.

The recall is calculated using Equation (2) in the previous section. In the case of Table 5, the true number of correct answers is two of the correct answers, "bedroom" and "bed", and the number of outputs of the system is two, "bedroom" and "bed". The above two equations were substituted into Equation (2), and the result was a recall of 1.

The F-value of 1 was obtained by substituting the obtained precision and recall into Equation (3). From this calculation result, it can be seen that the class can be successfully matched from the home ontology that is a candidate for guidance from the output of COMET-ATOMIC.

Table 7 shows an example of an unsuccessful speech, the system output, and the correct answer data. Table 8 shows the input of COMET-ATOMIC and the output results. In the case of the goodness-of-precision, as shown in Table 7, the number of correct answers in the system output is 0 because no candidate class label was found, and the number of outputs of the system is also 0. The above two numbers were substituted into Equation (1), and the result of the calculation was a precision of 0.

For the recall, in the case of Table 7, the number of true correct answers is one of the "fireplaces" in the correct answer data, and the number of outputs of the system is zero. The above two equations were substituted into Equation (2), and the result was a recall of 0. From this calculation result, we can see that the system did not output the guidance candidates well. As for the cause of the failure, Table 8 shows that the output results of the common-sense relationship between xNeed and xEffect indicate that the noun itself cannot be obtained well because of the class matching on the home ontology. From the output results of the common-sense relationship between AtLocation and xEffect, it can be seen that the nouns for class matching can be obtained, but the nouns "park", "car", "city", "cold", and "lake", which have nothing to do with the home environment, are inferred. If the nouns cannot be extracted from the output of COMET-ATOMIC, or if it outputs nouns that do not exist in the home ontology, it may not be able to output guidance candidates properly, resulting in failure. In order to obtain appropriate guidance candidates, it may be necessary to construct a home ontology that can be searched using not only nouns but also verbs and adjectives.

Table 5. Examples of correct utterances and system output.

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
I'm feeling really sleepy.	bedroom, bed	I will take you to your bed in your bedroom.	1.00	1.00	1.00

Table 6. COMET-ATOMIC input/output results and class matching results for the correct example speech.

Example of Speech	Commonsense Relation	COMET-ATOMIC Output Result	Class Matching Results with Home Ontology
I'm feeling really sleepy.	xNeed	"go to bed", "early sleep", "drink a lot"	bed
	xEffect	"falls asleep", "gets tired", "gets sleepy"	None
	AtLocation	"bed", "house", "hospital", "bedroom", "rest"	bedroom

Table 7. Examples of failed utterances and system output.

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
It's so cold, is not it?	Fireplace	I could not find a candidate for the guide.	0.00	0.00	0.00

Table 8. COMET-ATOMIC input/output results and class matching results for the failed example speech.

Example of Speech	Commonsense Relation	COMET-ATOMIC Output Result	Class Matching Results with Home Ontology
It's really cold, is not it?	xNeed	"to go outside.", "none"	None
	xEffect	"gets cold", "get" "wet"	None
	AtLocation	"park", "car", "city", "cold", "lake"	bedroom

5. Conclusions

5.1. Summary

In recent years, research has been conducted on Embodied AI, which uses agents such as virtual robots that learn in the real world or virtual space to perform various tasks in complex environments where various objects exist. However, many tasks can be solved by a one-way dialogue in which the virtual robot responds to a specific request 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 user and the virtual robot to have multiple conversations and infer the user's intentions.

In this study, we introduced knowledge-based reasoning to the study of Embodied AI and developed an interactive navigation system based on knowledge of the home environment and common-sense inference, in which the virtual robot can repeatedly interact with the user and infer the user's intentions and guide the user to the places in the virtual home environment that the user needs. We have developed a new interactive navigation system. In addition, we evaluated whether the proposed system was able to present appropriate locations as candidates based on the user's speech input about the home environment.

For the evaluation, 100 speech data extracted from the corpus of daily conversation, 100 speech data created by the subject, and the correct answer data for each data were created, and the precision, recall, and F value were calculated. As a result, we obtained a precision of 0.48, a recall of 0.58, and an F-value of 0.47 for the evaluation data extracted from the corpus of daily conversation and a precision of 0.48, a recall of 0.62, and an F-value of 0.49 for the evaluation data created by the subject. The results show that the system is able to provide users with somewhat appropriate guidance candidates.

5.2. Future Works

The first is the expansion of the natural language understanding dataset, which will enable us to extract appropriate intentions for ambiguous utterances by adding more examples of utterances with the intention of determining whether the utterance is routine and ambiguous. The second is the extension of the properties of home ontology, such as the time of day when objects in the home environment are frequently used. The third is the construction of home ontology about daily activities. In this study, only nouns are extracted from the output of COMET-ATOMIC and matched with home ontology in order to infer the content of speech from ambiguous speech. In the future, it will be possible to obtain more appropriate guidance candidates if we can construct home ontology about daily activities, as shown in [18,37], and infer.

In the future, it is expected that virtual robots learned in virtual space will be reflected in the real world and applied to the development of housekeeping robots, property guidance robots in real estate, and event venue guidance robots.

Author Contributions: Conceptualization, M.Y. and T.M.; methodology, A.S., M.Y., and T.M.; software, A.S.; investigation, A.S., M.Y., and T.M.; resources, A.S.; data curation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, M.Y. and T.M.; supervision, M.Y. and T.M.; project administration, T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Using the Corpus of Japanese Everyday Conversation (CEJC) [36], which collects speech data on daily conversation, we generated 100 speech examples for evaluation as input to our dialogue system using the procedure.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Examples of speech data and system output.

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
I'm hungry.	kitchen, fridge, pies, cereals, kitchen table, minced meat, cutlets, bananas, pudding, bread slices, cupcakes	There is a banana, a slice of bread, a cutlet, ground meat, cereal, pudding, plates, cupcakes, pie, refrigerator, and kitchen table in the kitchen. There are plates and cupcakes and a desk in the living room.	0.75	1.00	0.86
I need somewhere to sit	sofa, chair	There is a chair and a couch and a desk in the living room. There is a chair and a sofa and a kitchen table in the kitchen. There is a chair and a sofa in the bedroom.	0.75	1.00	0.86

Table A1. *Cont.*

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
I'm feeling a little sleepy.	bed, bedroom	There is a bed in the bedroom. There is a toilet in the bathroom.	0.50	1.00	0.67
It's so cold.	fireplace	No candidate for guidance found.	0.00	0.00	0.00
It's awfully bright	electric switch, dable lamp, wall lamp	I will take you to the bathroom in the bathroom.	0.00	0.00	0.00
I'm a little thirsty.	kitchen, refrigerator, milkshake	I will take you to the refrigerator in the kitchen.	1.00	0.67	0.80
I will read a book.	books	There are books and a TV in the living room. There are books in the bedroom. There is a TV in the kitchen.	0.50	1.00	0.67
My hair is in the way.	hair products	No candidate for directions found.	0.00	0.00	0.00
I will have to do a little cooking from now on	kitchen, refrigerator	I will take you to the refrigerator in the kitchen.	1.00	1.00	1.00
I'm gonna wipe my nose.	towels	There is a cupboard in the bedroom. There is a kitchen cabinet in the kitchen. There is a cabinet and a toilet in the bathroom.	0.00	0.00	0.00
I have a tummy ache.	toilet, painkiller	There is a cabinet and a toilet in the bathroom. There is a cupboard in the bedroom. There is a kitchen cabinet in the kitchen.	0.25	0.50	0.33
I just want to play.	no correct answer	I will take you to the bathroom in the bathroom.	0.00	0.00	0.00
I'd like to do some printing, too.	books, computer	There are books and a desk in the living room. There are books in the bedroom.	0.67	0.50	0.57
It's cold without a jacket.	closet	I take you to the bed in the bedroom.	0.00	0.00	0.00
Shall we sit down?	sofa,chair	There is a chair, a sofa, and a desk in the living room. There is a chair and a sofa and a kitchen table in the kitchen. There is a chair and a sofa in the bedroom.	0.75	1.00	0.86
I want to watch TV.	tv	There is a TV in the living room and in the kitchen room. Which would you like?	1.00	1.00	1.00
Do you have a kitchen?	kitchen	There is a banana, a slice of bread, a cutlet, ground meat, cereal, pudding, cupcakes and pie in the kitchen. There are cupcakes in the living room.	0.09	1.00	0.17
I do not have any pictures or anything.	picture frames, wall picture frames	There are books, wall picture frames and picture frames in the living room. There are wall picture frames and picture frames in the kitchen. There are books, wall picture frames and picture frames in the bedroom. There is a wall picture frame in the bathroom.	0.78	1.00	0.88
Maybe you could play the guitar for me.	no correct answer	No candidate for guidance found.	1.00	1.00	1.00
It's not very warm in here, is it?	fireplace	I will take you to the bathroom in the bathroom.	0.00	0.00	0.00

Table A1. *Cont.*

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
I will leave the door open.	door	Doors are in the kitchen, bedroom and bathroom room. Which one would you like?	1.00	1.00	1.00
I'm going to bed.	bed, bedroom	There is a bed in the bedroom. There is a toilet in the bathroom.	0.50	1.00	0.67
When you make something, you need tools.	no correct answer	I will take you to the box in the kitchen.	0.00	0.00	0.00
Where do you keep all your T-shirts?	y-shirt, closet drawer	There is a closet drawer and a Y-shirt in the living room. There is a closet drawer and a Y-shirt in the bedroom. There is a closet drawer, a towel rack, and a Y-shirt in the bathroom.	0.86	1.00	0.92
I was gonna put out the hand towels.	towel	There is a toilet in the bathroom. There is a desk in the living room.	0.00	0.00	0.00
Is there, like, a smoking room or something?	no correct answer	I will take you to the bathroom in the bathroom.	0.00	0.00	0.00
You'd have to light the stove or something.	fireplace	I will take you to the fireplace in the kitchen.	1.00	1.00	1.00
It's cold in here.	y-shirt, closet, pile of clothes	The closets are in the living room, the bedroom and the bathroom room. Which one would you like?	1.00	0.67	0.80
Do you have some kind of bowl?	bottles, mugs, coffee pot, cooking pot	There are bottles and mugs in the living room. There is a bottle, a coffee pot, a cooking pot and mugs in the kitchen. There is a bottle in the bathroom.	1.00	0.80	0.89
I like cold milk on a hot day.	refrigerator,milkshake	I will take you to the refrigerator in the kitchen.	1.00	0.50	0.67
I'm thirsty.	kitchen, refrigerator, milkshake	There is a toilet in the bathroom. There is a refrigerator in the kitchen.	0.50	0.67	0.57
Heat it up in the microwave.	microwave oven	There is a microwave, a banana, a slice of bread, a cutlet, ground meat, cereal, pudding, cupcakes, pie, and a refrigerator in the kitchen. There are cupcakes in the living room.	0.09	1.00	0.17
Where do you keep your T-shirts?	y-shirts, closet drawers	There is a closet drawer and a Y-shirt in the living room. There is a closet drawer and a y-shirt in the bedroom. There is a closet drawer, a towel rack, and a Y-shirt in the bathroom.	0.86	1.00	0.92
I feel awfully sleepy.	bed, bedroom	I will take you to the bed in the bedroom.	1.00	1.00	1.00
I'm kind of out of reach.	tv stand	There are books and a cell phone in the living room. There is a cell phone in the kitchen. There is a book in the bedroom.	0.00	0.00	0.00
I will call you in a bit.	cell phone	There is a book and a cell phone in the living room. There is a cell phone in the kitchen. There are books in the bedroom.	0.50	1.00	0.67

Table A1. *Cont.*

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
Let me take a picture.	picture frames, wall picture frames	There is a wall picture frame and a picture frame in the living room. There is a wall picture frame and a picture frame in the kitchen. There is a wall photo frame and a photo frame in the bedroom. There is a booth and a wall picture frame in the bathroom.	0.88	1.00	0.93
You want a banana?	banana	I will take you to the refrigerator in the kitchen.	0.00	0.00	0.00
Where is my phone?	portable	There are books and a cell phone in the living room. There is a cell phone in the kitchen. There are books in the bedroom. There is a booth in the bathroom.	0.40	1.00	0.57
It's in the way.	boxes, wall shelves, cupboards, bathroom shelves	Doors are in the kitchen, the bedroom and the bathroom room. Which one do you want?	0.00	0.00	0.00
Is it getting dark?	electric switch, dable lamp, wall lamp	I will take you to the bathroom in the bath room.	0.00	0.00	0.00
I have to go to the bathroom.	toilet	There is a booth, a toilet and a door in the bathroom. There is a door in the kitchen. There is a door to the bedroom.	0.20	1.00	0.33
Your hands are dirty.	toilet, sink, detergent	There is a closet and a toilet in the bathroom. There is a closet in the living room. There is a closet in the bedroom.	0.25	0.33	0.29
I will get you a glass of water.	refrigerator	I will take you to the refrigerator in the kitchen.	1.00	1.00	1.00
Wash the dishcloth properly, too.	sink, detergent	There is a closet drawer and a desk in the living room. There is a kitchen table in the kitchen. There is a closet drawer in the bedroom. There is a closet drawer in the bathroom.	0.00	0.00	0.00
Is there somewhere I can relax?	bedroom, living room, bathroom, bed, sofa	There is a desk and closet drawers in the living room. There is a closet drawer in the bedroom. There is a closet drawer in the bathroom.	0.43	0.60	0.50
Is there somewhere I can sit?	sofa, chair	There is a desk, sofa and chairs in the living room. There is a kitchen table, sofa and chairs in the kitchen. There is a sofa and chair in the bedroom.	0.75	1.00	0.86
I want to wash up	bathroom, sequins, towels	In the bathroom there is toilet paper, towels, toilet, and a sceen. There is detergent in the kitchen.	0.57	1.00	0.73
I want to take a bath	bathroom	I will take you to the toilet in the bathroom.	0.50	1.00	0.67
I need to cook	kitchen, bowls, cooking pot, refrigerator, kitchen shelves, minced meat, cutlet, banana, bread cutter	I will take you to the refrigerator in the kitchen.	1.00	0.22	0.36

Table A1. *Cont.*

Example of Speech	Correct Answer Data	System Output	Precision	Recall	F-Value
I need to get my period in order	cabinets, cupboards, kitchen cabinets, closets, piles of clothes, folders	There is a toilet and a cabinet in the bathroom. There is a cupboard in the bedroom. There is a kitchen cabinet in the kitchen.	0.75	0.50	0.60
I need to do my hair	toilet, hair products	I will take you to the toilet in the bathroom.	1.00	0.50	0.67
I want to read a book	books	The books are in the living room and the bedroom room. Which would you like?	1.00	1.00	1.00
I want to get out of this room	doors	There is a toilet and door in the bathroom. There is a door in the kitchen. There is a door and a bed in the bedroom.	0.60	1.00	0.75
I want to hang up my pictures	wall picture frames, picture frames	There is a wall picture frame and a book and picture frame in the living room. There is a wall picture frame and a picture frame in the kitchen. There is a wall picture frame, a book and a picture frame in the bedroom. There is a wall picture frame in the bathroom.	0.78	1.00	0.88
I'm getting sleepy.	bedroom, bed	There is a bed in the bedroom. There is a toilet in the bathroom.	0.50	1.00	0.67
I want to make coffee	coffee makers	I will take you to the fireplace in the kitchen.	0.00	0.00	0.00
Do you have anything to hang clothes on?	closets, hangers	There is a closet in the living room. There is a closet in the bedroom. There is a toilet and a closet in the bathroom.	0.75	0.50	0.60

References

1. Duan, J.; Yu, S.; Tan, H.L.; Zhu, H.; Tan, C. A Survey of Embodied AI: From Simulators to Research Tasks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, *6*, 230–244. [CrossRef]
2. Chaplot, D.S.; Gandhi, D.P.; Gupta, A.; Salakhutdinov, R.R. Object Goal Navigation using Goal-Oriented Semantic Exploration. In *Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 4247–4258.
3. Liu, X.; Muise, C. A Neural-Symbolic Approach for Object Navigation. In Proceedings of the 2nd Embodied AI Workshop (CVPR 2021), Virtual, 19–25 June 2021.
4. Ye, J.; Batra, D.; Das, A.; Wijmans, E. Auxiliary Tasks and Exploration Enable ObjectGoal Navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 16117–16126.
5. Ramrakhya, R.; Undersander, E.; Batra, D.; Das, A. Habitat-Web: Learning Embodied Object-Search Strategies from Human Demonstrations at Scale. In Proceedings of the CVPR, New Orleans, LA, USA, 19–24 June 2022.
6. Das, A.; Datta, S.; Gkioxari, G.; Lee, S.; Parikh, D.; Batra, D. Embodied Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1–10. [CrossRef]
7. Gordon, D.; Kembhavi, A.; Rastegari, M.; Redmon, J.; Fox, D.; Farhadi, A. IQA: Visual Question Answering in Interactive Environments. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4089–4098.
8. Yu, L.; Chen, X.; Gkioxari, G.; Bansal, M.; Berg, T.L.; Batra, D. Multi-Target Embodied Question Answering. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 6302–6311. [CrossRef]
9. Tan, S.; Xiang, W.; Liu, H.; Guo, D.; Sun, F. Multi-Agent Embodied Question Answering in Interactive Environments. In *Proceedings, Part XIII, Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 663–678._39. [CrossRef]

10. Zhu, F.; Zhu, Y.; Chang, X.; Liang, X. Vision-Language Navigation With Self-Supervised Auxiliary Reasoning Tasks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10009–10019. [CrossRef]
11. Thomason, J.; Murray, M.; Cakmak, M.; Zettlemoyer, L. Vision-and-Dialog Navigation. In *Proceedings of Machine Learning Research, Proceedings of the Conference on Robot Learning, Virtual, 16–18 November 2020*; Kaelbling, L.P., Kragic, D., Sugiura, K., Eds.; PMLR: Boulder, CO, USA, 2020; Volume 100, pp. 394–406.
12. Zhu, Y.; Zhu, F.; Zhan, Z.; Lin, B.; Jiao, J.; Chang, X.; Liang, X. Vision-Dialog Navigation by Exploring Cross-Modal Memory. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10727–10736. [CrossRef]
13. Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Gordon, D.; Zhu, Y.; Gupta, A.K.; Farhadi, A. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv* **2017**, arXiv:1712.05474.
14. Batra, D.; Gokaslan, A.; Kembhavi, A.; Maksymets, O.; Mottaghi, R.; Savva, M.; Toshev, A.; Wijmans, E. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv* **2020**, arXiv:2006.13171.
15. Wu, Y.; Wu, Y.; Gkioxari, G.; Tian, Y. Building generalizable agents with a realistic and rich 3D environment. *arXiv* **2018**, arXiv:1801.02209.
16. Pejsa, T.; Kantor, J.; Benko, H.; Ofek, E.; Wilson, A. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing; Association for Computing Machinery (CSCW '16), New York, NY, USA, 27 February–2 March 2016; pp. 1716–1725. [CrossRef]
17. Hwang, J.D.; Bhagavatula, C.; Bras, R.L.; Da, J.; Sakaguchi, K.; Bosselut, A.; Choi, Y. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In Proceedings of the AAAI, Virtual, 2–9 February 2021.
18. Vassiliades, A.; Bassiliades, N.; Gouidis, F.; Patkos, T. A Knowledge Retrieval Framework for Household Objects and Actions with External Knowledge. In *Proceedings of the Semantic Systems*; Blomqvist, E., Groth, P., de Boer, V., Pellegrini, T., Alam, M., Käfer, T., Kieseberg, P., Kirrane, S., Meroño-Peñuela, A., Pandit, H.J., Eds.; In the Era of Knowledge Graphs; Springer International Publishing: Cham, Switzerland, 2020; pp. 36–52.
19. Egami, S.; Nishimura, S.; Fukuda, K. A Framework for Constructing and Augmenting Knowledge Graphs using Virtual Space: Towards Analysis of Daily Activities. In Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, 1–3 November 2021; pp. 1226–1230. [CrossRef]
20. Zhang, Z.; Takanobu, R.; Zhu, Q.; Huang, M.; Zhu, X. Recent advances and challenges in task-oriented dialog systems. *Sci. China Technol. Sci.* **2020**, *63*, 2011–2027. [CrossRef]
21. Burtsev, M.; Seliverstov, A.; Airapetyan, R.; Arkhipov, M.; Baymurzina, D.; Bushkov, N.; Gureenkova, O.; Khakhulin, T.; Kuratov, Y.; Kuznetsov, D.; et al. DeepPavlov: Open-Source Library for Dialogue Systems. In *Proceedings of the ACL 2018, System Demonstrations*; Association for Computational Linguistics: Melbourne, PA, Australia, 2018; pp. 122–127. [CrossRef]
22. Ultes, S.; Rojas-Barahona, L.M.; Su, P.H.; Vandyke, D.; Kim, D.; Casanueva, I.; Budzianowski, P.; Mrkšić, N.; Wen, T.H.; Gašić, M.; et al. PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of the ACL 2017, System Demonstrations*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 73–78.
23. Chen, H.; Liu, X.; Yin, D.; Tang, J. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor. Newsl.* **2017**, *19*, 25–35. [CrossRef]
24. Puig, X.; Ra, K.; Boben, M.; Li, J.; Wang, T.; Fidler, S.; Torralba, A. VirtualHome: Simulating Household Activities Via Programs. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 8494–8502. [CrossRef]
25. Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. Habitat: A Platform for Embodied AI Research. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Long Beach, CA, USA, 15–20 June 2019.
26. Shen, B.; Xia, F.; Li, C.; Martín-Martín, R.; Fan, L.; Wang, G.; Pérez-D'Arpino, C.; Buch, S.; Srivastava, S.; Tchapmi, L.; et al. iGibson 1.0: A Simulation Environment for Interactive Tasks in Large Realistic Scenes. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 7520–7527. [CrossRef]
27. Beattie, C.; Leibo, J.Z.; Teplyashin, D.; Ward, T.; Wainwright, M.; Küttler, H.; Lefrancq, A.; Green, S.; Valdés, V.; Sadik, A.; et al. Stig DeepMind Lab. *arXiv* **2016**, arXiv.1612.03801.
28. Yan, C.; Misra, D.K.; Bennett, A.; Walsman, A.; Bisk, Y.; Artzi, Y. CHALET: Cornell House Agent Learning Environment. *CoRR* **2018**, arXiv:1801.07357.
29. Gao, X.; Gong, R.; Shu, T.; Xie, X.; Wang, S.; Zhu, S. VRKitchen: An Interactive 3D Virtual Environment for Task-oriented Learning. *arXiv* **2019**, arXiv:1903.05757.
30. Xiang, F.; Qin, Y.; Mo, K.; Xia, Y.; Zhu, H.; Liu, F.; Liu, M.; Jiang, H.; Yuan, Y.; Wang, H.; et al. SAPIEN: A SimulAted Part-Based Interactive ENvironment. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11094–11104. [CrossRef]
31. Gan, C.; Schwartz, J.; Alter, S.; Schrimpf, M.; Traer, J.; Freitas, J.D.; Kubilius, J.; Bhandwaldar, A.; Haber, N.; Sano, M.; et al. ThreeDWorld: A Platform for Interactive Multi-Modal Physical Simulation. *arXiv* **2020**, arXiv:2007.04954.

32. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open Source Language Understanding and Dialogue Management. *arXiv* **2017**, arXiv:1712.05181.
33. Morita, T.; Fukuta, N.; Izumi, N.; Yamaguchi, T. DODDLE-OWL: Interactive Domain Ontology Development with Open Source Software in Java. *IEICE Trans. Inf. Syst.* **2008**, *E91.D*, 945–958. [CrossRef]
34. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
35. Lamy, J.B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artif. Intell. Med.* **2017**, *80*, 11–28. [CrossRef] [PubMed]
36. Koiso, H.; Den, Y.; Iseki, Y.; Kashino, W.; Kawabata, Y.; Nishikawa, K.; Tanaka, Y.; Usuda, Y. Construction of the Corpus of Everyday Japanese Conversation: An interim report. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 4259–4264.
37. Zhu, Q.; Zhang, Z.; Fang, Y.; Li, X.; Takanobu, R.; Li, J.; Peng, B.; Gao, J.; Zhu, X.; Huang, M. ConvLab-2: An Open-Source Toolkit for Building, Evaluating, and Diagnosing Dialogue Systems. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 142–149. [CrossRef]