J-CRe3: A Japanese Conversation Dataset for Real-world Reference Resolution

Nobuhiro Ueda[1,2], **Hideko Habe**[2], **Yoko Matsui**[2], **Akishige Yuguchi**[3,2], **Seiya Kawano**[2,4], **Yasutomo Kawanishi**[2,4], **Sadao Kurohashi**[1,2,5], **Koichiro Yoshino**[2,4]
[1]Kyoto University, Kyoto, Japan, [2]Guardian Robot Project, R-IH, RIKEN, Kyoto, Japan,
[3]Tokyo University of Science, Tokyo, Japan, [4]Nara Institute of Science and Technology, Nara, Japan,
[5]National Institute of Informatics, Tokyo, Japan
{ueda,kuro}@nlp.ist.i.kyoto-u.ac.jp akishige.yuguchi@rs.tus.ac.jp
{hideko.habe,yoko.matsui,seiya.kawano,yasutomo.kawanishi,koichiro.yoshino}@riken.jp

Abstract

Understanding expressions that refer to the physical world is crucial for such human-assisting systems in the real world, as robots that must perform actions that are expected by users. In real-world reference resolution, a system must ground the verbal information that appears in user interactions to the visual information observed in egocentric views. To this end, we propose a multimodal reference resolution task and construct a Japanese Conversation dataset for Real-world Reference Resolution (J-CRe3). Our dataset contains egocentric video and dialogue audio of real-world conversations between two people acting as a master and an assistant robot at home. The dataset is annotated with crossmodal tags between phrases in the utterances and the object bounding boxes in the video frames. These tags include indirect reference relations, such as predicate-argument structures and bridging references as well as direct reference relations. We also constructed an experimental model and clarified the challenges in multimodal reference resolution tasks.

Keywords: Real-world Interaction, Reference Resolution, Phrase Grounding, Egocentric Video

1. Introduction

Human-assisting systems such as robots will be active in our living spaces in the near future. Such systems must understand the intention of users in the real world by grounding the referential expressions in language to real-world objects for cooperative action generation. Take the utterance, *Pour the coke here* as an example (Figure 1). For a robot to generate an appropriate action, the following arguments of predicate *pour* must be recognized: nominative: *robot*; accusative: *the coke*; and dative: *here*. Furthermore, it is crucial to ground entities (*the coke*), referential expressions (*here*), and even the agent of *pour* to their corresponding real-world entities.

Such reference resolution tasks with an egocentric view have been considered by existing works. Shirai et al. (2022) proposed a dataset to understand the cooking procedures by bridging the recipe texts and cooking videos. In an interactive scenario, SIMMC 2.1 (Kottur et al., 2021; Kottur and Moon, 2023) is a multimodal dialogue dataset that links referential expressions in dialogues with visual information in the virtual world.

In SIMMC 2.1, agents do not appear, and the movements or manipulation of objects are implemented as conceptual actions. This is because the dataset is oriented toward interaction in virtual space, although the physical relations between agent and objects remain important in a real-world interactions.

Another issue is that SIMMC 2.1 only focuses on limited reference relations; it has reference relation annotations only for phrases that appear in texts. However, referential phrases are frequently omitted in Japanese, called zero reference or zero anaphora (Sasano et al., 2008). For example, in the utterance, "こっちに持ってきて" (Can you bring (it) over here?[1]), the object to be brought is omitted.

We propose a multimodal reference resolution task that comprehensively handles zero references in real-world conversations involving object manipulation tasks. We constructed a dataset, J-CRe3: A Japanese Conversation dataset for Real-world Reference Resolution, which contains egocentric video and dialogue audio of real-world conversations between two people. The conversations involve a robot that is helping its master with daily mundane tasks, including many object manipulations. In addition, because the conversations are in Japanese, they naturally contain numerous zero references.

Figure 1 shows an example of our dataset. Each bounding box has a class name and an instance ID. Our dataset has two types of reference relations: textual and text-to-object reference relations. Textual reference relations include predicate-argument structures, bridging reference relations, and coreference relations. The text-to-object reference relation is a connection between a noun phrase and the bounding box to which it refers, as in the case of *here* and *sports drink* in the example. In addition, as this study handles zero references, we associate a predicate with the bounding boxes corresponding

[1]The indicator "(it)" is originally omitted in Japanese; it has been added to the English translation.

LREC-COLING 2024, pages 9489–9502
20-25 May, 2024.

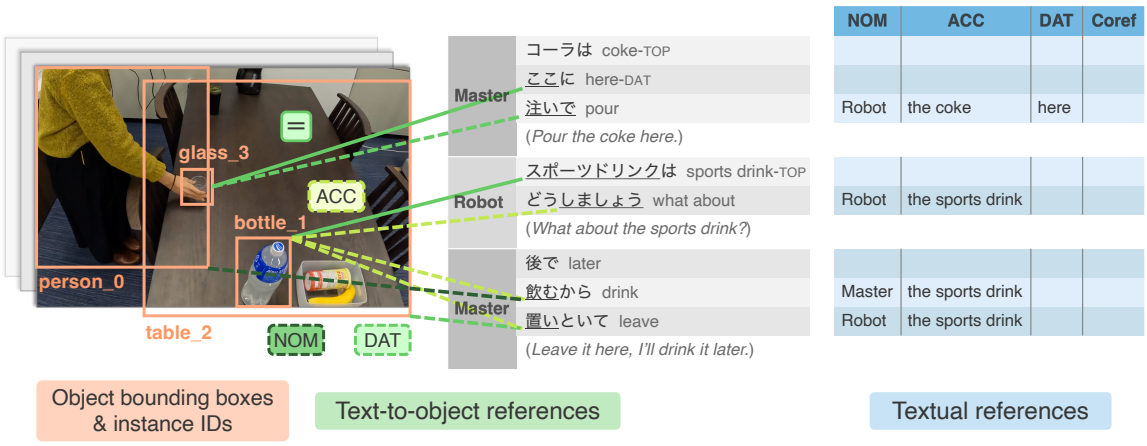

Figure 1: Example of J-CRe3. It has object bounding boxes (the orange rectangles), textual reference relations (the blue table), and text-to-object reference relations (the green lines). An object bounding box has a class name and an instance ID. Textual and text-to-object reference relations have 10–20 types of relations, including direct reference relations (=) and indirect reference relations corresponding to nominative (NOM), accusative (ACC), and dative (DAT) cases. For example, *sports drink* has a direct text-to-object reference relation (=) with the object bounding box, "bottle_1." Note that a particular case TOP shown in the example dialogue indicates an attached noun phrase is the sentence's topic.

to its arguments, even when those arguments are omitted. In Figure 1, the predicate *leave* and its dative argument exemplify zero reference. Although the dative argument of *leave* does not appear in the text,[2] *leave* and the object "table_2" are connected by the dative case relation, which the omitted argument would have with the predicate.

Our dataset consists of 93 videos and dialogue audio containing 2,131 utterances. The number of dialogues in our dataset is relatively small compared to the other related datasets (Table 1). However, our dataset has a fairly large number of unique images, in which all the objects referred to in the whole dialogue are densely annotated.

We also constructed an experimental model for our proposed multimodal reference resolution task to clarify the difficulty of the proposed task and the dataset. The proposed task can be divided into three widely studied tasks: textual reference resolution, object detection, and text-to-object reference resolution. Our experimental results showed that the accuracy of textual reference resolution was roughly the same as existing monologue datasets (F-scores of around 0.8). However, the text-to-object reference resolution task was demonstrated to be challenging (recall of around 0.4) and to have much room for improvement. Our dataset, including videos, audio, transcriptions, and annotations, is publicly available.[3] The source code and the weights of the resolution models used in this study are also publicly available.[4]

[2] *here* is omitted in the original Japanese text.

[3] https://github.com/riken-grp/J-CRe3

[4] https://github.com/riken-grp/multimodal-reference

2. Multimodal Reference Resolution

We propose a multimodal reference resolution task for real-world interactive systems that collaborate with humans. Given an egocentric image and the corresponding text as input, this task seeks the referents of nouns and predicates from objects in the image as well as phrases in the text. This task consists of three subtasks: textual reference resolution (Section 2.1), object detection (Section 2.2), and text-to-object reference resolution (Section 2.3).

2.1. Textual Reference Resolution

Textual reference resolution recognizes semantic relations among phrases in a text. Following previous studies (Ueda et al., 2020; Umakoshi et al., 2021; Ueda et al., 2023), we focus on predicate-argument structure (PAS), coreference, and bridging reference.

PAS is a set of relations between a predicate and its arguments, which correspond to *who* did/does *what* to *whom* for the predicate. Figure 1 (right) shows that the predicate *leave* has two arguments, *Robot* and *the sports drink*. Note that here, the accusative case *it* for the predicate *leave* is omitted, and instead, *the sports drink* is marked as the accusative case. Because the omitted argument *it* refers to *the sports drink*, this is an example of zero reference, and thus PAS is used to annotate zero references (Hangyo et al., 2012).

Coreference is a phenomenon where two (or more) noun phrases refer to the same entity in the

Dataset	# Annotated images	Text type	# Dialogues	Video	Zero reference
RefCOCO (Yu et al., 2016)	20k	Referring expression	-	✗	✗
RefCOCO+ (Yu et al., 2016)	142k	Referring expression	-	✗	✗
RefCOCOg (Mao et al., 2016)	26k	Referring expression	-	✗	✗
VisualGenome (Krishna et al., 2017)	108k	Caption	-	✗	✗
Flickr30k Entities (Plummer et al., 2017)	30k	Caption	-	✗	✗
VisCoref (Yu et al., 2019)	5k	Dialogue	5,000	✗	✗
Visual Recipe Flow (Shirai et al., 2022)	6k	Cooking recipe	-	✗	✗
BioVL2 (Nishimura et al., 2021, 2022)	3k	Experimental procedure	-	✓	✗
EPIC-KITCHENS (Damen et al., 2022)	277k	Narration	-	✓	✗
RefEgo (Kurita et al., 2023)	226k	Referring expression	-	✓	✗
SIMMC 2.1 (Kottur and Moon, 2023)	2k	Dialogue	11,244	✗	✗
J-CRe3 (ours)	11k	Dialogue	93	✓	✓

Table 1: Comparison of image or egocentric video datasets with relations between phrases and objects.

real world. A bridging reference relation is an indirect relation between two noun phrases where one noun phrase (anaphor) refers to the other (antecedent) and the latter complements the essential meaning of the former. These semantic relations are all crucial for dialogue understanding by interactive robots operating in the real world.

2.2. Object Detection

Object detection identifies and locates objects within an image. The output of this task is object bounding boxes, as shown in Figure 1 (left). The detected bounding boxes are fed to the text-to-object reference resolution task.

2.3. Text-to-object Reference Resolution

Text-to-object reference resolution identifies the referents of nouns and predicates, similar to textual reference resolution. Yet, the referents are selected from the object detection's output. In Figure 1, this task predicts the edges between the words and the object bounding boxes. The task of detecting an object directly referenced by a phrase is known as phrase grounding (Kamath et al., 2021; Gupta et al., 2020) or referring expression comprehension (Qiao et al., 2020). However, in real-world conversations, where the interlocutors share the visual information and phrases are frequently omitted, resolving direct references is insufficient for adequate comprehension of their utterances. Thus, we also consider indirect relations, including PAS and bridging references, which involve zero references.

3. J-CRe3 Dataset

We constructed the J-CRe3 dataset for multimodal reference resolution. It consists of video and audio recordings of real-world conversation scenes between two people, audio transcriptions, and annotations of various reference relations. Through the lens of applications to human-assisting systems, we assumed that the two interlocutors are a master and an assistant robot and prepared three dialogue locations: a living room, a dining room, and a kitchen.

In this section, we describe the construction procedure and statistics of J-CRe3. First, we collected dialogue scenarios through crowdsourcing (Section 3.1). Then, we recruited actors for the master and robot roles and recorded the egocentric videos and the dialogue audio of their conversations following the collected scenarios (Section 3.2). Finally, we labeled the bounding boxes and the reference relations to the audio transcriptions and the video frames per second (Section 3.3). The statistics of the dataset are described in Section 3.4.

3.1. Dialogue Scenario Collection

We collected a variety of realistic dialogue scenarios through crowdsourcing. In the crowdsourcing task, the workers were shown pictures of the room and objects to be used in the conversation recording.[5] The workers then wrote dialogue texts along with the interlocutors' actions and surrounding situations. 101 workers participated in our task, and 180 scenarios were collected. The number of utterances per scenario was limited to 10–16 to ensure that the dialogues were not too long and had sufficient context. We manually filtered out the collected scenarios that lacked feasibility, a sufficient number of referential expressions, and sufficient descriptive granularity for the situation and manually modified the remaining scenarios for more naturalness. Appendix A shows an example of a modified scenario.

3.2. Conversation Recording

We recruited five actors and paired two of them to perform the master and robot roles and recorded

[5]The crowdsourcing interface is shown in Appendix C.

their in-person conversations following the modified scenarios. The recording was conducted in a laboratory furnished to resemble a living room, a dining room, and a kitchen.[6] Both actors were equipped with close-talking microphones to record their speeches. The actor playing the robot had a head-mounted RGB camera[7] to capture an egocentric video during the conversation. We installed four fixed RGB cameras in the corner of the ceiling in the laboratory to record third-person videos, *i.e.,* the entire room's overview.

3.3. Annotation

We annotated the collected conversational audio and egocentric videos for multimodal reference resolution. The third-person videos were used only as a reference of the objects that are occluded or out of view in the egocentric videos.

We transcribed every utterance[8] from the conversational audio and converted the egocentric videos into image sequences by extracting the frames every second. We annotated each utterance with timestamps at its beginning and ending points to ensure proper alignment with the videos.

The following sections describe the annotation for the textual reference resolution, the object detection, and the text-to-object reference resolution.

3.3.1. Textual Reference Annotation

We annotated the transcribed conversational text with predicate-argument structures, coreference relations, and bridging reference relations. We followed the annotation guidelines of an existing textual reference corpus (Hangyo et al., 2012). Although the existing corpus contains monologue-style written text, our dataset contains dialogue-style spoken text. Therefore, we defined additional guidelines for colloquial expressions, including casual replies.[9]

In Figure 1, coreference relation would be annotated between *the sports drink* and *it*, if *it* were mentioned in the dialogue. For bridging reference, if Robot said, "Drink it early because the expiration date is approaching," the relation would be annotated between *the sports drink* and *the expiration date*.

3.3.2. Bounding Box Annotation

We annotated each image extracted from the egocentric videos with object bounding boxes. We also labeled an object class name and an instance ID for each bounding box. The class name was selected from a set of 1,203 classes defined in the LVIS dataset (Gupta et al., 2019), which is widely used in object detection tasks. An instance ID is an identifier that uniquely distinguishes each object and must be assigned consistently throughout the video.

We used a general object detector called Detic (Zhou et al., 2022) and a multi-object tracker called StrongSORT (Du et al., 2023) to automatically annotate the bounding boxes, which were then manually corrected.

3.3.3. Text-To-Object Reference Annotation

We assigned reference relations to every combination of phrases in the text and bounding boxes in the image. The phrases include noun phrases and predicates. For noun phrases, we assigned bounding boxes to objects directly referenced or with bridging reference relations. For predicates, we assigned bounding boxes to objects corresponding to the predicates' arguments. Text-to-object reference annotation is similar to the textual reference annotation where the reference target is extended to object bounding boxes.

Text-to-object reference annotation involves a significantly large number of relations to be annotated. However, we can eliminate most of them by utilizing previously assigned instance IDs. For example, given that a reference relation is assigned to a bounding box of a "cup" in a specific video frame, the reference relations of the same "cup" in other frames can be automatically assigned. We can also utilize previously assigned textual reference relations. For example, consider a case where the following text reference annotation has already been assigned to an utterance text:

(1) Can you put that cup on the table?
(NOM: robot, ACC: cup, LOC: the table)

Once we assign a direct reference relation between *cup* and a bounding box, the relation between the predicate *put* and the bounding box can be automatically labeled as the accusative case. Therefore, no accusative text-to-object annotation for *put* is required. The same applies to *the table*.

A characteristic phenomenon in real-world conversation is the use of demonstrative pronouns that refer to such locations, as *here* and *there*. Identifying such indicated locations is essential for robots

[6] The furnished recording area is a room of approximately 6.5m x 6.8m in which the three locations are set up without walls.

[7] We used GoPro HERO10 Black. It has a wide-angle lens and can capture the whole room from the corner.

[8] An *utterance* is almost always a series of sentences made by a speaker before a turn transition. However, following Yoshino et al. (2018), when an utterance's pause exceeds 500 msec and the utterance's semantic content is complete at that point, an utterance is delimited.

[9] https://github.com/riken-grp/J-CRe3/blob/main/docs/annotation_guideline.pdf

Figure 2: Example of regional bounding box annotation

collaborating with humans. In this study, we describe such phrases as regional referring expressions and assign corresponding regions to them in text-to-object reference annotation. Unlike object bounding boxes, the bounding boxes of these regions are not uniquely determined, and their position and size depend on an annotator's subjective judgment. For this reason, we assigned a special class name *region* to these bounding boxes to distinguish them from other object bounding boxes. Figure 2 shows an example of regional bounding box annotation for the following utterance:

(2) Can you put that cup away right there?
(=: region_11)

As expression *right there* does not refer to any object but rather to a part of the table, the region is tagged as a regional bounding box. The regional bounding box and *right there* are assigned a direct reference relation.

3.4. Statistics

Table 2 shows the statistics of our dataset. The number of object instances and classes indicates that our dataset contains bounding boxes of diverse objects. The number of unique object classes in our entire dataset is 166. The number of zero references is significantly larger than that of direct references, suggesting the importance of resolving zero references.

To quantify the textual diversity of our dataset, we calculated a dataset-level distinct-1 and distinct-2 scores (Li et al., 2016), where we counted unique n-grams over the entire dataset. These scores of the dialogue texts in our dataset were 0.087 and 0.336, respectively, while those of the dialogue texts in SIMMC 2.1 were 0.054 and 0.285, indicating that our scenarios are more diverse than SIMMC 2.1.[10]

[10]We used the Japanese morphological analyzer, Juman++ (Tolmachev et al., 2018) and the nltk toolkit (Bird et al., 2009) to tokenize Japanese and English texts, respectively. For SIMMC 2.1 dataset, we randomly sampled utterances from the dev split to match the number of words in our dataset for fair comparison.

	Train	Val.	Test	Total
Recorded Data				
Dialogues	75	9	9	93
Utterances	1,746	155	230	2,131
Sentences	2,176	197	279	2,652
Morphemes	13,780	1,319	1,720	16,819
Total duration (sec)	8,747	919	1,358	11,024
Textual Reference Annotation				
Predicates	2,780	264	342	3,386
Nominative args.	2,824	274	349	3,447
Accusative args.	1,138	94	135	1,367
Dative args.	1,399	133	146	1,678
Nominative-2 args.	527	63	62	652
Bridging anaphors	463	30	50	543
Coref. mentions	1,094	104	121	1,319
Bounding Box Annotation				
Frames	8,780	922	1,360	11,062
Bounding boxes	65,431	5,079	9,184	79,694
Object instances	1,435	119	187	1,741
Object classes	158	42	49	-
Text-To-Object Reference Annotation				
Direct reference	1,306	102	160	1,568
Nominative case	2,091	155	297	2,543
Accusative case	1,134	74	136	1,344
Dative case	1,228	96	148	1,472
Nominative-2 case	522	47	65	634
Bridging references	440	18	41	499
Zero references	6,016	453	708	7,177

Table 2: Statistics of J-CRe3.

4. Reference Resolution

We clarified the difficulty of the multimodal reference resolution task and the constructed dataset by training and evaluating an experimental model on it. Multimodal reference resolution consists of three subtasks: textual reference resolution, object detection, and text-to-object reference resolution. In our experiments, we independently trained and evaluated the textual reference resolution and the other two tasks and then combined the results.

4.1. Textual Reference Resolution

4.1.1. Task Settings

Textual reference resolution consists of predicate-argument structure (PAS) analysis, bridging reference resolution, and coreference resolution. Following Ueda et al. (2020, 2023), we extracted predicates and eventive noun phrases for PAS analysis, non-eventive noun phrases for bridging reference resolution, and noun phrases for coreference resolution from the transcribed utterances.

Unlike bridging reference and coreference resolutions, PAS analysis classifies the relations between predicates and their arguments into a set of predefined labels, *i.e.,* cases. In this study, we focus

Task		J-CRe3 Endophora	J-CRe3 Exophora	J-CRe3 All	KWDLC Endophora	KWDLC Exophora	KWDLC All
PAS analysis	Nominative	0.84 (210)	0.86 (211)	0.85 (422)	0.87 (3649)	0.77 (2269)	0.83 (5918)
	Accusative	0.87 (164)	0.06 (3)	0.85 (167)	0.86 (2218)	0.45 (238)	0.82 (2456)
	Dative	0.89 (74)	0.81 (118)	0.84 (192)	0.80 (1239)	0.64 (582)	0.75 (1822)
	Nominative-2	0.00 (6)	0.89 (76)	0.85 (82)	0.58 (179)	0.57 (104)	0.58 (283)
Bridging reference resolution		0.83 (56)	0.41 (8)	0.79 (64)	0.69 (1897)	0.55 (208)	0.68 (2106)
Coreference resolution		0.72 (71)	0.60 (15)	0.70 (86)	0.84 (1746)	0.77 (350)	0.82 (2097)

Table 3: F-scores of textual reference resolution: Endophora is a reference to entities that appear in the text. Exophora is a reference to entities that do not. We fine-tuned our model with three different random seeds and report the mean performances. Numbers of gold references are shown in parentheses.

on four cases: nominative, accusative, dative, and nominative-2.[11]

4.1.2. Method

We employed a word selection model for textual reference resolution following previous works (Ueda et al., 2020, 2023). This model formulates all three tasks as a word selection task in which the model selects a word from a given text as the referent of a target word (e.g., predicate). We added two layers of feed-forward neural networks for each task on top of a pre-trained encoder model[12] and fine-tuned the whole model.

For fine-tuning the model, we utilized the following textual reference resolution corpora: the Kyoto University Text Corpus (Kurohashi and Nagao, 1998; Kawahara et al., 2002),[13] the Kyoto University Web Document Leads Corpus (KWDLC, Hangyo et al., 2012),[14] the Annotated Fuman Kaitori Center Corpus,[15] and the Wikipedia Annotated Corpus.[16] These corpora contain 9,207 documents, an amount that mitigates the data scarcity problem of our dataset. We mixed the training split of our dataset with the existing corpora to train the model.

To combine these corpora with our dataset, we converted its speaker labels ("master" and "robot") to be *relative*. Existing corpora provide textual reference relations not only between phrases but also between a phrase and an entity that does not appear in the text, such as "writer" and "reader." Such references are called exophoras. To utilize these labels, we transformed the exophora labels "master" and "robot" in our dataset into "speaker" or "listener" for each utterance. We then treated these relative labels as equivalent to the "writer" and "reader" labels, allowing us to mix all the datasets. Although the model is trained to predict the relative labels, they can be converted to absolute labels ("master" and "robot") using the speaker labels assigned to each utterance.[17]

4.1.3. Results

Table 3 shows the experimental results. In all the tasks, we achieved comparable results to those of KWDLC, a monologue web corpus widely used for Japanese textual reference resolution. A notable difference between J-CRe3 and KWDLC is the scores of exophora reference resolution in the nominative, dative, and nominative-2 cases. Arguments of the nominative and nominative-2 cases generally include an agent; thus, they are easier to resolve when the agent is a master or a robot. Also, arguments of the dative case often include a listener who is asked to do something, making the resolution easier due to the limited referents. Therefore, further verification is needed to evaluate the performance when the number of interlocutors exceeds two.

4.2. Object Detection and Text-To-Object Reference Resolution

Text-to-object reference resolution can be divided into direct reference resolution (denoted as "=" in Figure 1) and indirect reference resolution (denoted as "NOM", "ACC", and "DAT"). The direct reference resolution is also called phrase grounding. Although a phrase grounding model cannot resolve indirect references, including zero references, it is

[11] Nominative-2 is used for a common Japanese construction in which a predicate has two nominative arguments.

[12] We used a DeBERTa V2 model pre-trained on Japanese corpora (https://huggingface.co/ku-nlp/deberta-v2-large-japanese).

[13] https://github.com/ku-nlp/KyotoCorpus

[14] https://github.com/ku-nlp/KWDLC

[15] https://github.com/ku-nlp/AnnotatedFKCCorpus

[16] https://github.com/ku-nlp/WikipediaAnnotatedCorpus

[17] We also examined a model using absolute labels of "master" and "robot," although the relative labeling method achieved better scores in most cases.

actively studied and many models and datasets have been proposed (Kamath et al., 2021; Gupta et al., 2020; Plummer et al., 2017; Nakayama et al., 2020). To investigate the extent to which these models can solve text-to-object reference resolution, even partially, this section discusses experiments with an existing phrase grounding model.

4.2.1. Task Settings

Given an image and a corresponding text description, phrase grounding detects the objects in the former that correspond to each phrase in the latter (Kamath et al., 2021; Gupta et al., 2020). J-CRe3 consists of image sequences from videos, not individual images. That is, each phrase in a text (i.e., a transcribed utterance) has multiple images for grounding. For simplicity, we considered the utterance containing the phrase and limited the grounding target to the video frames between the utterance's start and the next one's start.

As an evaluation metric, we used Recall@k, which is the major evaluation metric for phrase grounding models. It measures whether a model can rank the "correct" box among its top k predictions. A box is considered correct if the Intersection-over-Union (IoU) between the predicted and the ground-truth boxes exceeds a predetermined threshold. We set the threshold to 0.5, following Kamath et al. (2021).

4.2.2. Method

To address the scarcity of training data, we fine-tuned a pre-trained phrase grounding model using existing phrase grounding datasets. As MDETR (Kamath et al., 2021), which serves as the base phrase grounding model, performs object detection and phrase grounding in an end-to-end manner, we do not need a separate object detector.

However, a pre-trained MDETR model has two issues. The first is language mismatch. The MDETR model is trained on datasets for English phrase grounding and referring expression comprehension. Therefore, we fine-tuned the model using the Flickr30k Entities JP dataset (Nakayama et al., 2020), which is a Japanese translation of the Flickr30k Entities dataset (Plummer et al., 2017), a commonly used phrase grounding dataset.

The second issue is domain mismatch. Although the MDETR model is trained on photos intentionally taken by a human photographer, the images in our dataset are frames of egocentric video, which often contain blurred or occluded objects. We performed additional fine-tuning on our dataset to address this issue.

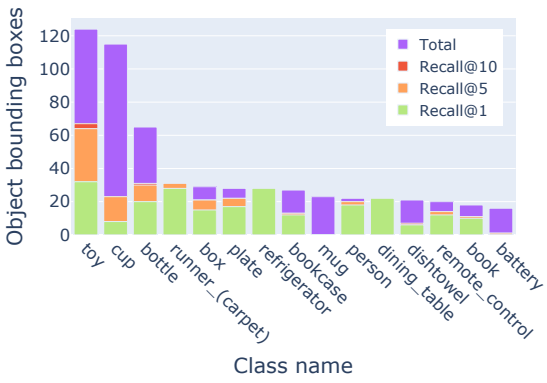

Figure 3: Distribution of Recall@k and number of objects for each object class: Figure shows top 15 classes. Although we show the difference between Recall@5 and Recall@10 in red in the figure, there were very few of them.

4.2.3. Training Details

In the first stage of fine-tuning, the text encoder of MDETR, which is a RoBERTa base (Liu et al., 2019), was replaced with a multilingual encoder, an XLM-RoBERTa base (Conneau et al., 2020). To adapt the entire MDETR model to the new encoder, we fine-tuned it with a mixture of RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016), Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), and Flickr30k Entities JP.[18] As English is the dominant language in the training data, we froze the text encoder except for the final layer to prevent overfitting to English texts.

In the second stage, we first converted our dataset into the same format as the Flickr30k Entities, where each image has multiple text descriptions of 10–20 words. We divided the dialogue text into segments of two consecutive utterances and treated each one as a single text description. Note that we removed segments without any references to objects. Other training details are described in Appendix B.

4.2.4. Results

Table 4 shows the Recall@k of the phrase grounding model. The first-stage fine-tuning contributes to both the J-CRe3 and Flickr30k Entities JP performances, suggesting that it largely mitigates the language mismatch issue. The second-stage fine-tuning also improved the J-CRe3 performance. However, compared to Flickr30k Entities JP, the

[18] In our preliminary experiment, combining all the datasets improved the performance more than using only Flickr30k Entities JP.

Model	J-CRe3			Flickr30k Entities JP		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
MDETR (ENB3)	0.007 (3/407)	0.012 (5)	0.012 (5)	0.007	0.010	0.010
+ FT1	0.214 (87/407)	0.381 (155)	0.403 (164)	**0.671**	**0.819**	**0.845**
+ FT1 + FT2	**0.337** (137/407)	**0.474** (193)	**0.511** (208)	0.222	0.293	0.309
MDETR (ENB5)	0.005 (2/407)	0.007 (3)	0.007 (3)	0.018	0.023	0.024
+ FT1	0.231 (94/407)	0.376 (153)	0.420 (171)	**0.675**	**0.818**	**0.845**
+ FT1 + FT2	**0.410** (167/407)	**0.494** (201)	**0.504** (205)	0.215	0.260	0.265

Table 4: Performances of phrase grounding models: *ENB3* and *ENB5* indicate EfficientNet-B3 and EfficientNet-B5, which are used as backbone of MDETR. *FT1* denotes model is fine-tuned on the existing datasets: RefCOCO, RefCOCO+, RefCOCOg, Visual Genome, GQA, and Flickr30k Entities JP. *FT2* denotes model is further fine-tuned on J-CRe3. Values in parentheses are correctly resolved references.

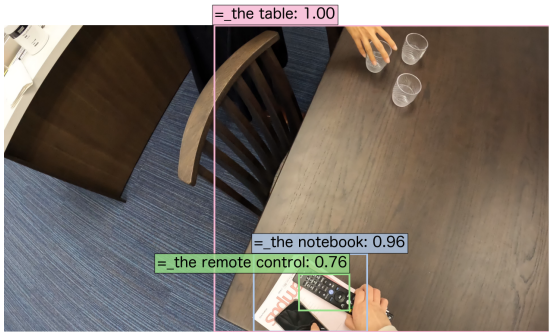

Figure 4: Example of phrase grounding for utterance "The notebook and mobile phone should be put on the table in the next room, and the remote control should be put on the sofa, right?" (translated). Predicted phrases and their confidences are shown with colored object bounding boxes. The confidence threshold is 0.5.

scores are significantly lower, which suggests that domain mismatch remains a significant issue.

Figure 3 visualizes the distribution of Recall@k for each object class. The classes with low recall include *cup*, *mug*, and *battery*. One possible reason is that objects in these classes are relatively small and often appear with similar objects in the same class. Generally, detecting small objects is known to be a challenging task (Rekavandi et al., 2023). Using dedicated object detectors instead of MDETR possibly improves the coverage of detected objects and the performance of phrase grounding.

Figure 4 shows an example of the phrase grounding results on our dataset. When we focus on such object names as *the notebook*, *the remote control*, and *the table*, the bounding boxes seem correctly identified. However, *the table* is referred to as *the table in the next room*, which is not shown in the frame. Although we trained the model not to ground phrases whose corresponding objects did not appear in the frame, the model grounded *the table* with high confidence. Another problem is the failure to ground the phrase *mobile phone* to

Temporal loc.	Recall@1	Recall@5	Recall@10
First half	0.361 (110/305)	0.472 (144)	0.489 (149)
Second half	**0.366** (120/328)	**0.500** (164)	**0.512** (168)
After	0.313 (60/192)	0.396 (76)	0.417 (80)

Table 5: Relation between the temporal location of the video frames and Recall@k. We evaluated 18 dialogues, including the validation and test sets.

the smartphone beside the remote control. Indeed, identifying a smartphone from this image alone is challenging due to insufficient visual information. Requiring a high level of textual and visual context understanding is one of the factors that make phrase grounding on our dataset challenging.

4.2.5. Impact of the Behavior of the Robot Actor on the Model Performance

The conversations in our dataset are intended to be between a robot and a human; however, in its recording, human actors are employed to play the roles of robots. Thus, human thoughts interfere in the actor's behaviors, such as gaze shift and object grasping, potentially resulting in egocentric videos that are conveniently analyzable for the model. For instance, in a scene with multiple cups, if the master says, "Fill the biggest cup with water," the robot actor will look at and pick up the biggest cup. After the actor picks up the cup, the video frame is expected to prominently feature that cup, making it easier for the model to identify the "biggest cup" as the referent.

To investigate the impact of the actor's behavior on the model performance, we evaluated the phrase grounding model with respect to the temporal location of the frames corresponding to the target utterance. If the model uses the actor's behaviors as clues, its performance is better on later frames as the actor's actions are responses to the master's utterances. We classified the frames into three categories based on their temporal locations and evaluated Recall@k for each category.

Task	Recall@1	Recall@5	Recall@10
Nominative reference resolution	0.064 (79/1230)	0.070 (86)	0.072 (88)
Accusative reference resolution	0.199 (67/336)	0.232 (78)	0.235 (79)
Dative reference resolution	0.035 (25/719)	0.047 (34)	0.047 (34)
Nominative-2 reference resolution	0.000 (0/399)	0.000 (0)	0.000 (0)
Bridging reference resolution	0.198 (17/86)	0.198 (17)	0.198 (17)
Direct reference resolution	0.410 (167/407)	0.494 (201)	0.504 (205)

Table 6: Result of text-to-object reference resolution: Numbers in parentheses denote correctly resolved references and their total number.

- Frames temporally corresponding the first half of the target utterance (**First half**)
- Frames temporally corresponding the second half of the target utterance (**Second half**)
- Frames between the end of the target utterance and the start of the next utterance (**After**)

Table 5 shows the results. The performance is higher for the Second half frames than the First half ones, suggesting that the actor's behavior serves as a clue for the model. Therefore, when evaluating a system using our dataset, it is important to focus on the performance for the earlier frames corresponding to the target utterance.

4.3. Combining the Results

The phrase grounding model cannot handle indirect reference relations. In other words, it cannot identify an object that corresponds to a predicate's arguments, nor can it identify an object that has a bridging reference relation with a noun. However, we can resolve these references by combining the output of the phrase grounding model with that of the textual reference resolution model.

Table 6 shows the combined results of the object detection and text-to-object reference resolution. The performance on all tasks is significantly low due to error propagation issues. In particular, the Recall@1 of the phrase grounding model is 0.410, which is the upper bound of the performance and greatly impacts the results.

5. Related Work

Egocentric video datasets that involve human-object manipulation tasks include Ego4D (Grauman et al., 2022), EPIC-Kitchens (Damen et al., 2022), Home Action Genome (Rai et al., 2021), and BioVL2 (Nishimura et al., 2021, 2022). Ego4D, EPIC-Kitchens, and Home Action Genome feature everyday actions similar to our datasets; BioVL2 contains experimental videos from the biochemistry field. These datasets focus on actions in videos and do not address the localization of objects based on dialogue. Although BioVL2 provides bounding boxes, the target objects are restricted to those in the experimental protocol and in contact with the experimenter's hand.

The tasks that associate text-to-object bounding boxes include referring expression comprehension (REC, Yu et al., 2016) and phrase grounding (Plummer et al., 2017). REC detects object regions that corresponds to a given text description. Phrase grounding is more challenging, since it involves detecting object regions for all the phrases within a text description. However, both tasks only address object regions directly related to the text. In contrast, multimodal reference resolution encompasses indirect relations, such as bridging reference relations and predicate-argument structures.

6. Conclusion

We proposed a multimodal reference resolution task to realize a robot that interacts and collaborates with humans in the real world. This task consists of textual reference resolution, which identifies the phrases that are referred to in texts; object detection, which detects referent object candidates in images; and text-to-object reference resolution, which identifies referent objects from the output of object detection. To train and evaluate models for this task, we constructed a Japanese Conversation dataset for Real-world Reference Resolution (J-CRe3). It is based on real-world conversations and is expected to contribute to the development of more practical dialogue understanding systems.

Our future plans include improving the resolution model for this task. Although the system used in our experiments analyzed textual and text-to-object references independently, resolving these relations in an integrated manner could improve the performance.

Another way to improve the resolution model is to expand the size and domain of the dataset. We can use text or image generation models to expand the dataset at a low cost. Specifically, it is possible to append generated dialogues to images in existing phrase grounding datasets.

Limitations

Constructing a real-world conversation dataset is costly, and our dataset is limited in size, making it insufficient for independently training a model. However, this issue can be mitigated by combining it with other datasets or pre-trained models.

Our dataset is designed for interactions between a master and a domestic assistant robot. Consequently, the model's generalizability to other contexts remains uncertain, such as outdoor environments, or different types of human-robot interactions, such as interactions with multiple robots.

Acknowledgements

This work was supported by Kyoto University Science and Technology Innovation Creation Fellowship (Information / AI field). This work was partially supported by JSPS KAKENHI Grant Number 22H03654.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)*, 130:33–55.

Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25:8725–8737.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Kottur, et al. 2022. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *Computer Vision – ECCV 2020*, pages 752–768, Cham. Springer International Publishing.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 535–544.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr–modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese Relevance-tagged Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Satwik Kottur and Seungwhan Moon. 2023. Overview of situated and interactive multimodal conversations (SIMMC) 2.1 track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 235–241, Prague, Czech Republic. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73.

Shuhei Kurita, Naoki Katsura, and Eri Onami. 2023. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15214–15224.

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese Parsed Corpus while Improving the Parsing System. In *International Conference on Language Resources and Evaluation (LREC'98)*, pages 719–724.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

Taichi Nishimura, Kojiro Sakoda, Atsushi Hashimoto, Yoshitaka Ushiku, Natsuko Tanaka, Fumihito Ono, Hirotaka Kameko, and Shinsuke Mori. 2021. Egocentric Biochemical Video-and-Language Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3129–3133.

Taichi Nishimura, Kojiro Sakoda, Atsushi Ushiku, Atsushi Hashimoto, Natsuko Okuda, Fumihito Ono, Hirotaka Kameko, and Shinsuke Mori. 2022. BioVL2: An Egocentric Biochemical Video-and-Language Dataset. *Journal of Natural Language Processing*, 29(4):1106–1137. In Japanese.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision (IJCV)*, 123(1):74–93.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. 2021. Home Action Genome: Cooperative Compositional Action Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11184–11193.

Aref Miri Rekavandi, Shima Rashidi, Farid Boussaid, Stephen Hoefs, Emre Akbas, and Mohammed bennamoun. 2023. Transformers in small object detection: A benchmark and survey of state-of-the-art.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 769–776, Manchester, UK. Coling 2008 Organizing Committee.

Keisuke Shirai, Atsushi Hashimoto, Taichi Nishimura, Hirotaka Kameko, Shuhei Kurita, Yoshitaka Ushiku, and Shinsuke Mori. 2022. Visual Recipe Flow: A Dataset for Learning Visual State Changes of Objects with Recipe Flows. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3570–3577, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2023. KWJA: A Unified Japanese Analyzer Based on Foundation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 538–548, Toronto, Canada. Association for Computational Linguistics.

Masato Umakoshi, Yugo Murawaki, and Sadao Kurohashi. 2021. Japanese zero anaphora resolution can benefit from parallel texts through neural transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1920–1934, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. 2018. Japanese Dialogue Corpus of Information Navigation and Attentive Listening Annotated with Extended ISO-24617-2 Dialogue Act Tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2922–2927, Miyazaki, Japan. European Language Resources Association (ELRA).

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision – ECCV 2022*, pages 350–368, Cham. Springer Nature Switzerland.

A. Scenario Example

Table 7 shows an example of our collected scenarios. The text in the parentheses describes the scene contexts, which were used only for conversation recording and were not included in our dataset. Texts outside the parentheses contain such referential expressions as *over there* and *it*, making it difficult to understand the dialogue text without visual contexts.

Speaker	**Utterance**
Master	I want to pack this doll. Can you shred some paper and make some cushioning material for it?
Robot	Yes, I can. What paper should I use?
Master	Use that paper (points at a pile of magazines in the room's corner). Take enough to fill that cardboard box.
Robot	(Shreds old magazines) Is this enough?
Master	(Checks inside the cardboard box) Yeah, that's plenty. Can you bring me that bubble wrap over there?
Robot	Yes, of course (Hands it over) Should I wrap it up as well?
Master	No, it's too fragile, let me do it. (Wraps the doll) Bring me some tape.
Robot	Where is it?
Master	(Pointing to a shelf) I think it's probably in the cupboard on the right. (Places the doll in the box)
Robot	(Goes to the shelf and grabs some wrapping tape) Is this okay?
Master	Yes. I've packed the dolls, so tape the lid and carried it to the front door.
Robot	I understand (The box is taped shut). I'm done, so I'll leave the box at the front door.

Table 7: Example of collected scenarios: Text in parentheses is scene context and was not used as dialogue.

B. Training Details

We trained the phrase grounding model with the hyper-parameters shown in Table 8, following Kamath et al. (2021). The EfficientNet-B3 and EfficientNet-B5 models were downloaded from https://github.com/ashkamath/mdetr?tab=readme-ov-file#pre-training. The XLM-RoBERTa base model was downloaded from https://huggingface.co/xlm-roberta-base.

Settings	FT1	FT2
Detection Backbone	EfficientNet-B3 or EfficientNet-B5	
Text Encoder	XLM-RoBERTa base	
Batch size	8	
Training Epochs	2	
Learning Rate	1e-4	
Learning Rate (detection backbone)	5e-5	
Learning Rate (text encoder)	5e-5	
Weight Decay	1e-4	
Gradient Clipping	0.1	
Exponential Moving Average Decay	0.9998	

Table 8: Hyper-parameters used for fine-tuning phrase grounding model.

C. Crowdsourcing Interface

Figure 5 shows the crowdsourcing interface used for the scenario collection. Note that the interface shown is an English translation of the original Japanese interface.

CrowdWorks

Job to write conversation scenarios between a living room assistant robot and the owner

Job Details

【 Summary 】
This job involves thinking of dialogue scenarios between a living room assistant robot and its owner.

【Request Content】
Task:
In the near future, a time will come when robots and humans live together in homes, with robots assisting in various tasks.
For this task, please imagine scenarios where the owner in the living room converses with the robot while the robot performs tasks.
Ensure the dialogue between the owner and the robot consists of **at least 5 exchanges** each, and design it so that the conversation naturally concludes.
Also, please provide a brief summary of the scenario initially.

Conditions for the dialogue scenario:
- A photo of the living room with the owner and the robot is attached. Please consider interactions **including elements shown in the photo**.
- Below is a list of tasks the robot can perform. Create scenarios with the assumption that you are conversing with a robot capable of these tasks.
- Generally, it's just you and the robot in the room. If there are scenarios involving other people, like serving tea to a guest, please specify in the summary.

Tasks the robot can perform:
- Handing over and receiving
- Carrying, moving, transferring, delivering
- Picking up, extracting, removing
- Putting back, storing, inserting
- Tidying up, throwing away
- Aligning, arranging, organizing
- Picking up, grabbing, lifting
- Teaching, conveying
- Distributing, serving (drinks)
- Pouring (drinks)
- Brewing (tea, coffee)
- Playing (movies, music)
- Turning on, operating (TV, radio)
- Reading, reading aloud (books)
- Putting (children) to bed

The envisioned robot:
- It's a domestic assistant robot. It will do everything you ask to the best of its abilities.
- It can perform tasks equivalent to humans, such as fetching items or washing dishes. Conversely, it cannot do things difficult for humans, such as lifting heavy objects.
- It possesses intelligence equivalent to humans and can interpret vague expressions like "that" or "this" based on the context.

Example response (directions in parentheses):
Summary: The robot fetches the newspaper and glasses
Owner: I'd like to read the newspaper, could you get it for me from over there?
Robot: Understood. (Goes to get the newspaper and finds two) Which one would you like?
Owner: Um, maybe the one on the table.
Robot: This one, right? Here you go.
Owner: Thank you. Oh, this is yesterday's. Not this one, could you bring the one underneath?
Robot: Understood. (Goes to get the other newspaper) Here it is.
Owner: This is it. By the way, I've been having trouble reading small print recently because my eyesight is getting worse.
Robot: Should I bring your glasses as well?
Owner: Yes. Um, where did I put them?
Robot: They are in front of the TV, I'll bring them right away.

【 Reward 】
For each dialogue scenario between the owner and the robot you think of, you will receive 100 yen (excluding tax).

【Notes】
If the content is clearly mechanical or unnatural, or if there are any omissions, it may be subject to rejection.
If you have any other questions, please feel free to contact us.
We look forward to your application!

Attached Files


living1.png
17.5MB


living2.png
17MB

Figure 5: Translated crowdsourcing interface used for collection of scenarios in living room