Symbol Grounding from Natural Conversation for Human-Robot Communication

Ye Kyaw Thu
Okayama Prefectural University
Okayama Prefecture, Japan
ye@c.oka-pu.ac.jp

Takuya Ishida
Okayama Prefectural University
Okayama Prefecture, Japan
t.ishida2706@gmail.com

Naoto Iwahashi
Okayama Prefectural University
Okayama Prefecture, Japan
iwahashi@c.oka-pu.ac.jp

Tomoaki Nakamura
The University of Electro-Communications, Tokyo
Tokyo, Japan
naka_t@apple.ee.uec.ac.jp

Takayuki Nagai
The University of Electro-Communications, Tokyo
Tokyo, Japan
tnagai@ee.uec.ac.jp

ABSTRACT
UPDATED—September 4, 2017. This paper proposes a new approach for research on chat-like conversational systems that enable robots to acquire physically grounded knowledge through natural interaction with humans. The proposed approach combines research on chat-like conversational systems, language acquisition, and symbol grounding in order to realize physically situated and natural human-robot interaction. In contrast to previous approaches for chat-like conversation, the proposed approach focuses on utterances which are situated in physical environments surrounding humans and robots. Based on the proposed approach, we develop a concrete method that enables robots to learn object image concepts and the words describe them from object-teaching utterances made by humans. The method is composed of two processes: (1) the detection of object-teaching utterances from chat-like conversation and (2) the learning of object image concepts and the words describing them. It applies a linear support vector machine, multimodal hierarchical Dirichlet process, and term frequency-inverse document frequency process. The experimental results show that the method enabled robots to learn object image concepts and the words that describe them through multimodal chat-like interactions with humans.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; See http://acm.org/about/class/1998/ for the full list of ACM classifiers. This section is required.

Author Keywords
Chat-like conversational system, Physically grounded knowledge, Support Vector Machine (SVM), Multimodal Hierarchical Dirichlet Process (MHDP), Term frequency-inverse document frequency (tf-idf)

HAI '17, October 17–20, 2017, Bielefeld, Germany
© 2017 ACM. ISBN 978-1-4503-5113-3/17/10…$15.00
DOI: https://doi.org/10.1145/3125739.3132611

INTRODUCTION
Recently, research on chat-like conversational systems has been thriving (e.g., [11]). Several chat-like conversational robots have been developed, such as Palro, Pepper. These studies have focused on utterances involving topics that are not related to things in front of both humans and robots.

Research on language acquisition and symbol grounding for robots has been also thriving [2], [23]. These studies have focused on the acquisition of physically grounded knowledge through utterances that express physical entities, such as objects and motions, in front of both humans and robots. Hereafter, we refer this kind of utterances as grounded utterances. Most previous studies have focused on learning physically grounded knowledge without any prior symbolic knowledge.

Despite this bipolarization trend in research, the problem of how to acquire physically grounded knowledge based on grounded utterances through natural interaction has yet to be fully explored. Indeed, existing chat-like conversational systems are unable to learn the correspondence between language and entities in front of humans and robots. For example, even when a human says, "You know, this is a remote control", while showing the TV remote control to a robot, the robot cannot learn the correspondence between the word "remote control" and the physical object of the TV remote control. In order for robots to understand human grounded utterances correctly, they should process integrated information on the utterances and physical environments. We have to consider the possibility that there exists both grounded and ungrounded utterances in chat-like conversations. The treatment of grounded utterances may be an important topic for research on chat-like conversational systems. This kind of research approach must be explored.

Based on such a research approach, we aim to develop robots that are capable of learning the correspondence between language and physical objects through natural, chat-based interactions with humans. In this paper, we focus on object-teaching utterances as grounded utterances. We propose a method that can detect object-teaching utterances from chat-like conversations between a human and a robot, and can

learn the concepts of physical objects and the words that describe them.

RELATED WORK

Generally, early works on situated referential grounding have focused on computational models that connect linguistic referring expressions to realize the environment [9], [8], [21], [18]. These proposals were manual or automatic approaches such as "parsing semantic information", "mapping procedures" and "building functions between visual features and words" for a situated referential grounding system. Edmonds (1994) [7], Heeman and Hirst (1995) [10] proposed a symbolic reasoning based approach for collaborative dialogue. However, pure symbolic approaches were insufficient for situated grounding. Hence, a hybrid approach which combined symbolic reasoning and machine learning for interpreting referential grounding dialogue was proposed by DeVault and Stone (2009) [5]. However, their setting environment was a simplistic block world and to mediate perceptual differences Liu (2014) [17] proposed probabilistic labeling for referential grounding. Although probabilistic labeling significantly outperforms the state-space search, the grounding performance is still rather poor even for the top-3 hypotheses [17]. All these mentioned approaches unable to learn a new physically grounded knowledge.

PROPOSED METHOD

Overview

The setup of this study is as follows. A participant and a robot engage in a chat-like conversation across a table with several objects on it (see Figure 1). The chat-like conversation is realized with a rule-based (e.g., [26]) chat-like conversational system that we developed [28]. During chat-like conversation, the participant makes both ungrounded and object-teaching utterances, while holding or pointing at the objects. By repeating such interactions, the set of pairs of speech information obtained by a microphone and visual information obtained by an RGB-D sensor, is recorded as data to be used by our proposed method.

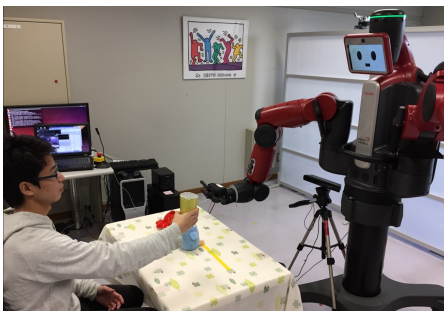

Figure 1. An example environment of conversation between human and robot

An overview of the proposed system is shown in Figure 2. The method uses the abovementioned recorded data as input, and generates the set of pairs composed of an object image concept and the word describing it. Two main processes are involved. The first is the detection of object-teaching utterances. The system classifies human utterances as either grounded utterance or object-teaching utterance using speech and behavioral information. The second process is the learning of object image concepts and words. The system makes two inferences simultaneously: (1) what object in each scene that each utterance expresses; and (2) what word in the utterance describes the object. The details of the method are described in the following subsections.

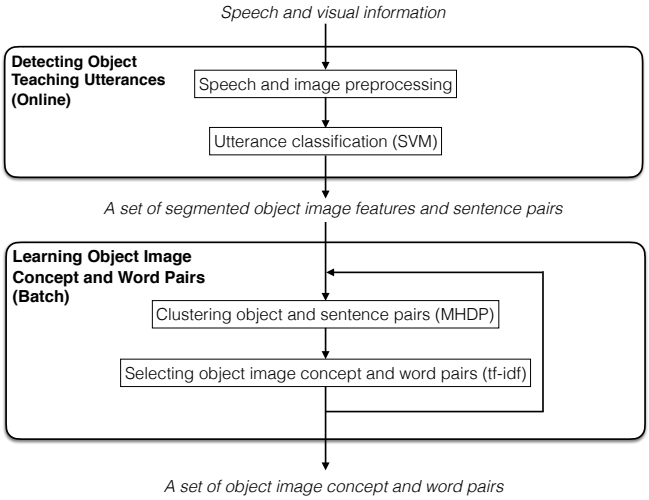

Figure 2. Overview of proposed method

Detecting object-teaching utterances

This process makes the decision on whether the input utterance is an object-teaching or an ungrounded utterance via an online process, and prepares the next process of learning object image concept and word pairs with the set of the object image features and object-teaching utterance sentence pairs. In this data, each sentence can correspond to multiple object images. This process includes speech and image preprocessing and utterance classification.

Speech and image preprocessing

For speech processing, we used rospeex [15], a cloud-based multilingual communication package for Robot Operating System. Here, we used rospeex for speech recognition in Japanese. In addition, morphological analysis was performed on speech recognition results with the morphological analyzer software MeCab [22].

For image processing, images as visual information were captured using the RGB-D sensor, Microsoft Kinet v1. The objects in the images were detected, and each object image was segmented with in-house software L-Core [13], [12]. Feature extraction from the segmented object images was implemented using a convolutional neural network (CNN) approach [16], [6]. This approach used the CAFFE deep learning framework [14] with IMAGENET, an open trained image network model [4]. Behavioral information on whether the human was holding or pointing at an object was extracted with L-Core.

Utterance classification

This process classifies each utterance as either ungrounded utterance or object-teaching utterance using morphological analysis information and human behavioral information, which were obtained by the speech and image preprocessing. The linear support vector machine (SVM) software LIBSVM [3] was used for the classification. Figure 3 shows the overall classification process of ungrounded utterance or not with SVM. As features, we used grasping and pointing information of human motion, morphological information of utterance and also question words information of previous sentence.

Learning object image concept and word pairs

This process learns object image concept and descriptive word pairs in a batch way using the output from the first process of the detection of object-teaching utterances. Note that

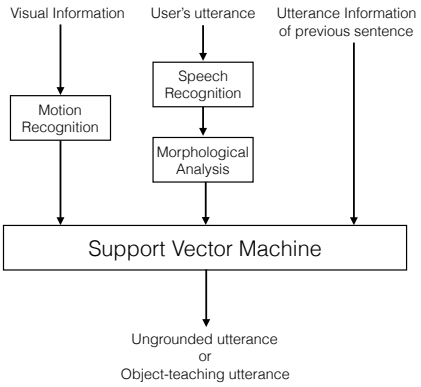

Figure 3. Classification of ungrounded utterance or object-teaching utterance with SVM

the detected utterances can contain the ungrounded utterances owing to detection error. This process includes the process for clustering object and sentence pairs and the process for selecting object image concept and word pairs. These processes are run in a loop.

Clustering object and sentence pairs

This process carries out the clustering of object image features and descriptive sentence pairs, and learns object concepts using a multimodal hierarchical Dirichlet process (MHDP) [19][25]. MHDP is an extension of the hierarchical Dirichlet process [24]. Figure 4 shows the graphical model of MHDP. This model represents each object category as multinomial distributions of the object image features and words in sentences. Here, x^v_{ij} and x^w_{ij} represent the j-th object image and word information of the i-th pair of a segmented object image features and a sentence, respectively. The information of modality $m \in \{v, w\}$ is generated from a multinomial distribution whose parameter is θ^m_k. The parameter θ^m_k is drawn from a Dirichlet distribution parameterized by α^m_k for category k. As the result of clustering, object category ID is given to each pair of a segmented object image and a sentence.

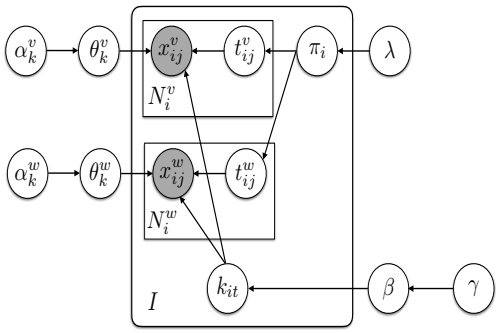

Figure 4. Graphical representation of MHDP model

Selecting object image concept and word pairs

The input data of this process are the results of the clustering process with MHDP. In this data, each sentence can correspond to multiple object images and include multiple words. The process selects an appropriate object image concept and descriptive word pair among all possible combinations of object image concepts and words in each object category, by maximizing the value of term frequency-inverse document frequency (tf-idf) [27]. The tf-idf weight is a statistical weight measure often used in information retrieval and text mining [20]. Generally, it is used to evaluate the importance of a word to a document in a collection or corpus. Here, we assumed that the category classification output of the MHDP model (i.e., sentence level) was one document. Hence, the calculation of the tf-idf weight is as follows:

$$tf\text{-}idf_{w,k} = tf_{w,k} \times idf_w$$

$$tf_{w,k} = \frac{\text{No. of times word } w \text{ appears in the object category k}}{\text{Total no. of words in the object category k}}$$

$$idf_w = log \frac{\text{Total no. of object categories}}{\text{No. of object categories with word } w}$$

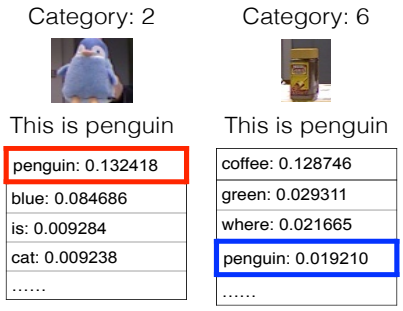

Figure 5. An example of two object categories

As a final result, the highest tf-idf value of word and object category pair will be selected. For example, if we have two object categories (Category 2 and Category 6) with tf-idf values for words as shown in Figure 5, the selected result will be the object Category 2 with the highest tf-idf value (0.132418) word penguin pair. The process of word and object pair selection with tf-idf weight can be expressed as follow:

$$(\hat{w}, \hat{k}) = \arg\max_{w,k} tf\text{-}idf_{w,k}$$

EXPERIMENTS

We conducted experiments to evaluate the proposed "detecting object teaching utterances" and "learning object image concept and word pairs" processes.

Detecting object teaching utterance

Data

We used 2,000 utterances from chat-like human-robot conversations [28] together with visual information. Only 200 utterances (10% of the corpus) were object teaching utterance. Some examples of dialogs conducted in the experiment can be seen in Table 1. Here, italic sentences are object-teaching utterances. We used 10 objects for this experiment: an accessory box, a coffee bottle, lunch box, stuffed toy penguin, two black stuffed toy cats (small and big), two stuffed toy fishes (red and yellow), and two cups (red and yellow), as shown in Figure 6. An example of image file that we used can be seen in Figure 7. Object detection from images have to be done for extracting features with CNN. There were seven classes of images. The word recognition rate with the rospeex speech recognition engine was 70.8%. Three types of experimental conditions for utterance classification with SVM are shown in Table 2. Here, condition-1 is using only morphological analysis information, condition-2 is using only behavioural information and condition-3 is using both of them.

Result

The precision, recall, and F-measure of utterance classification with the SVM model, and 10-fold cross validation are

Figure 6. The ten objects used in the experiment

Figure 7. An example of image that we used in the experiment

shown in Table 3. Among the three experimental conditions (see Table 2), Condition 3 (using both audio and visual information) resulted the highest F-measure value of 87.9 %.

Learning object image concept and word pairs

Data

The input data was the output of the previous process of detecting object teaching utterances (see Figure 2). We used 196 utterances that were selected as object-teaching utterance by the linear SVM model and 417 related object image pairs. The proportion of correctly recognized words that described objects was 35.7%.

Result

We evaluated the process of learning object image concept and word pairs with P_w, P_c and P_{wc}. The explanation for each probability value is as follows:

P_w: probability of selecting correct word in each sentence

P_c: probability of selecting correct object image concept for each sentence

P_{wc}: probability of selecting both correct word and object image concept for each sentence

By analyzing the results of symbol grounding using the without loop and with loop methods, we found that the with loop approach yielded better probability values for all P_w, P_c and P_{wc} compared to the without loop approach (see Table 4).

DISCUSSION

Although the first process for detecting object-teaching utterances provided high accuracy, it still needs improvement. Accuracy can be improved by increasing the word recognition rate from the current rate of 70.8 %. Similarly, the second process for learning object image concept and word pairs was affected by the use of training data acquired by the speech recognition engine.

Table 1. Some examples of dialogue conducted in the experiment

Human	Robot
Do you know any toys?	I am not familiar with toy.
Here is the stuffed toy.	Oh, I see.
Do you like animals?	I like dogs.
I like this penguin.	I got it.

Table 2. Experimental conditions for utterance classification with linear SVM

Condition	1	2	3
Morphological Analysis Info	○	–	○
Behavioural Info	–	○	○

Table 3. Result of detecting object-teaching utterance with linear SVM

Condition	Precision	Recall	F-measure
Condition-1	84.9%	78.7%	81.4%
Condition-2	23.9%	20.2%	21.7%
Condition-3	89.0%	87.4%	87.9%

Table 4. Results of learning object image concept and word pairs without loop (MHDP ⇒ tf-idf) and with loop (MHDP ⇒ tf-idf ⇒ MHDP ⇒ tf-idf)

Method	P_w	P_c	P_{wc}
w/o loop	31% (61/196)	30% (59/196)	10% (19/196)
w/ loop	35% (69/196)	57% (112/196)	28% (54/196)

Because the use of the behavioral information could improve the performance of the first process, we can expect the improvement of the second process using similar information. In addition, although the first process did not use the information on word order in sentences, it may be useful. In order to improve the first process, the information on word order in sentences may be useful. We plan to develop the method using conditional random field. Currently, the second process runs in a batch way. We plan to convert it into an online process using online MHDP [1] in the near future.

CONCLUSION

This paper proposed a new approach for chat-like conversational systems that enable robots to acquire physically grounded knowledge through natural interaction with humans. The approach combined a chat-like conversational system with aspects of language acquisition and symbol grounding to realize physically situated natural human-robot interaction. We developed a concrete method that enabled robots to learn object image concepts and the associated descriptive words based on object-teaching utterances by humans. The method combined SVM, MHDP, and tf-idf process effectively. Experimental results showed the validity of the method clearly. This is a very important step that bring us closer to the goal of creating a framework for natural human-robot communication for real-word tasks.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI (grant number 15K00244) and JST CREST ("Symbol Emergence in Robotics for Future Human-Machine Collaboration")

REFERENCES

1. Takaya Araki, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano, and Naoto Iwahashi. 2012. Online Object Categorization Using Multimodal Information Autonomously Acquired by a Mobile

Robot. *Advanced Robotics* 26, 17 (2012), 1995–2020. DOI:http://dx.doi.org/10.1080/01691864.2012.728693

2. Angelo Cangelosi and Matthew Schlesinger. 2014. *Developmental Robotics: From Babies to Robots*. The MIT Press.
3. Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages. DOI:http://dx.doi.org/10.1145/1961189.1961199
4. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
5. David DeVault and Matthew Stone. 2009. Learning to Interpret Utterances Using Dialogue History. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 184–192. http://dl.acm.org/citation.cfm?id=1609067.1609087
6. Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *CoRR* abs/1310.1531 (2013). http://arxiv.org/abs/1310.1531
7. Philip G. Edmonds. 1994. Collaboration on Reference to Objects That Are Not Mutually Known. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2 (COLING '94)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1118–1122. DOI: http://dx.doi.org/10.3115/991250.991333
8. Peter Gorniak and Deb Roy. 2007. Situated Language Understanding as Filtering Perceived Affordances. *Cognitive Science* 31, 2 (2007), 197–231. DOI: http://dx.doi.org/10.1080/15326900701221199
9. Peter Gorniak and Deb Roy. 2011. Grounded Semantic Composition for Visual Scenes. *CoRR* abs/1107.0031 (2011). http://arxiv.org/abs/1107.0031
10. Peter A. Heeman and Graeme Hirst. 1995. *Collaborating on Referring Expressions*. Technical Report. Rochester, NY, USA.
11. Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015. Towards Taxonomy of Errors in Chat-oriented Dialogue Systems. In *SIGDIAL Conference*.
12. Naoto Iwahashi. 2007. Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations. *Human Robot Interaction* (September 2007), 95–118.
13. Naoto Iwahashi, Komei Sugiura, Ryo Taguchi, Takayuki Nagai, and Tadahiro Taniguchi. 2010. Robots that Learn to Communicate: A Developmental Approach to Personally and Physically Situated Human-Robot Conversations. In *Dialog with Robots, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*. http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2320
14. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 675–678. DOI: http://dx.doi.org/10.1145/2647868.2654889
15. Sugiura Komei, Hori Chiori, and Zettsu Koji. 2013. rospeex: A Cloud-based Spoken Language Communication Toolkit for ROS. . *CNR,* 113, 248 (oct 2013), 7–10. http://ci.nii.ac.jp/naid/110009784563/en/
16. Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*. 2278–2324.
17. Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014. *Probabilistic labeling for efficient referential grounding based on collaborative discourse*. Vol. 2. Association for Computational Linguistics (ACL), 13–18.
18. C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. *ArXiv e-prints* (June 2012).
19. T. Nakamura, T. Nagai, and N. Iwahashi. 2011. Multimodal categorization by hierarchical dirichlet process. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1520–1525. DOI: http://dx.doi.org/10.1109/IROS.2011.6094763
20. Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended Boolean Information Retrieval. *Commun. ACM* 26, 11 (Nov. 1983), 1022–1036. DOI: http://dx.doi.org/10.1145/182.358466
21. Alexander Siebert and David Schlangen. 2008. A Simple Method for Resolution of Definite Reference in a Shared Visual Context. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue (SIGdial '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 84–87. http://dl.acm.org/citation.cfm?id=1622064.1622080
22. Kudo Taku. 2005. MeCab : Yet Another Part-of-Speech and Morphological Analyzer. *http://mecab.sourceforge.net/* (2005). http://ci.nii.ac.jp/naid/10019716933/en/
23. Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata, and Hideki Asoh. 2015. Symbol Emergence in Robotics: A Survey. *CoRR* abs/1509.08973 (2015). http://arxiv.org/abs/1509.08973
24. Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* 101 (2004).
25. Nakamura Tomoaki, Araki Takaya, Nagai Takayuki, and Iwahashi Naoto. 2013. Multimodal Object Categorization Based on Hierarchical Dirichlet Process by a Robot. *Transactions of the Society of Instrument and Control Engineers* 49, 4 (apr 2013), 469–478. DOI: http://dx.doi.org/10.9746/sicetr.49.469
26. Richard S. Wallace. 2009. *The Anatomy of A.L.I.C.E.* Springer Netherlands, Dordrecht, 181–210.
27. Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting TF-IDF Term Weights As Making Relevance Decisions. *ACM Trans. Inf. Syst.* 26, 3, Article 13 (June 2008), 37 pages. DOI: http://dx.doi.org/10.1145/1361684.1361686
28. Kazuma Yamamoto, Takuya Ishida, Naoto Iwahashi, Ye Kyaw Thu, and Takeo Kunishima. 2016. Symbol Grounding in Human-Robot Dialogs using Visual and Linguistic Cues. In *Proceedings of the Human-Agent Interaction Symposium 2016*. the Human-Agent Interaction Symposium Program Committee, Tokyo, Japan, P–36. http://hai-conference.net/symp2016/proceedings/pdf/P-36.pdf