Mixed Reference Interpretation in Multi-turn Conversation

Nanase Otake, Shoya Matsumori, Yosuke Fukuchi, Yusuke Takimoto and Michita Imai
Faculty of Science and Technology, Keio University, Japan

Keywords: Conversation Context, Demonstrative Word Interpretation, Contextual Reference, Situational Reference, Mixed Reference Interpretation.

Abstract: Contextual reference refers to the mention of matters or topics that appear in the conversation, and situational reference to the mention of objects or events that exist around the conversation participants. In conventional utterance processing, the system deals with either contextual or situational reference in a dialogue. However, in order to achieve meaningful communication between people and the system in the real world, the system needs to consider Mixed Reference Interpretation (MRI) problem, that is, handling both types of reference in an integrated manner. In this paper, we propose DICONS, a method that sequentially estimates an interpretation of utterances from interpretation candidates derived from both contextual reference and situational reference in a dialogue. In an experiment in which DICONS handled this task with both contextual and situational references, we found that it could properly judge which type of reference had occurred. We also found that the referent of the demonstrative word in each context and situation could be properly estimated.

1 INTRODUCTION

In human conversation, both contextual reference and situational reference are intermingled. Contextual reference is to refer matters or topics that has mentioned in the conversation. Situational reference, on the other hand, is to mention objects or events that exist around the conversation participants. These two types of reference are illustrated in Fig. 1, which shows example scenarios in which a person and a robot are shopping at a supermarket. In the scene on the left, where the person asks "Which one is better for dinner, fish or meat?" and the robot answers "Fish", contextual reference occurs because the robot selects the word "fish" as the answer to the question based on the content of the human utterance. In the scene on the right, where fish and meat are displayed on the shelf and the person standing in front of it asks "Which one is better?" and the robot answers "Fish", situational reference occurs because the robot answers the question based on the information about objects in the environment. As shown in this example, in real-world dialogue, both contextual reference and situational reference may occur in the same situation.

A common problem with previous methods is that they deal with either contextual or situational reference in dialogue, not both. In other words, the system is designed assuming a situation where only context reference occurs, as represented by a chatbot, or only situational reference, as represented by an instruction task to the robot. However, in order to deal with more realistic situations in the real world, a system needs to handle the Mixed Reference Interpretation (MRI) problem, that is, both types of reference coexist, and we cannot determine in advance which one will occur. Another problem is that previous methods do not deal with the "sequentiality of dialogue", where the context gradually develops according to the utterance of each speaker. In order to enable a dialog system and a person to dynamically understand each other's utter-

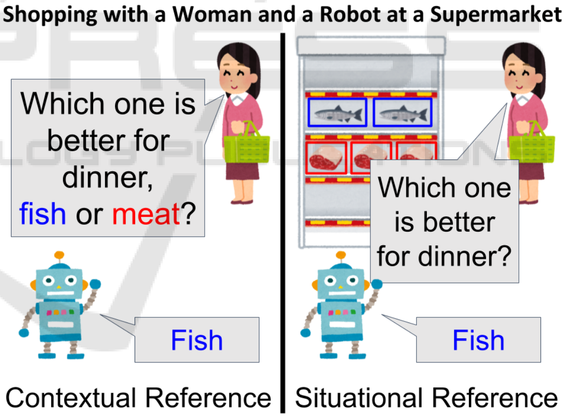
Shopping with a Woman and a Robot at a Supermarket
Which one is better for dinner, fish or meat?
Fish
Which one is better for dinner?
Fish
Contextual Reference
Situational Reference

Figure 1: Example of contextual reference and situational reference. We call the task in which speakers need to deal with both references Mixed Reference Interpretation (MRI).

Otake, N., Matsumori, S., Fukuchi, Y., Takimoto, Y. and Imai, M.
Mixed Reference Interpretation in Multi-turn Conversation.
DOI: 10.5220/0010260103210328
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 321-328
ISBN: 978-989-758-484-8

ances, it is necessary, in ambiguous states in which contextual reference or situational reference occur, to decide which one it is by estimating sequentially which reference the current utterance corresponds to.

In this paper, we focus on the MRI problem in multi-turn conversation processing. We propose Dynamic and Incremental Interpretation of Contextual and Situational References in Conversational Dialogues (DICONS), a method that sequentially estimates an interpretation of utterances from interpretation candidates derived from contextual reference and situational reference in a dialogue. DICONS simultaneously performs a probabilistic search of interpretation candidates considering the contextual reference and considering the situational reference in a multi-turn conversation, compares each interpretation, and then estimates which type of reference it is. In this way, even in a conversation where the reference type is not clear, DICONS can estimate the reference type and the referent.

This paper is structured as follows. In Chapter 2, we provide an overview of previous works on contextual and situational references and describe SCAIN (Takimoto et al., 2020), which is the base algorithm of DICONS. Chapter 3 presents DICONS in detail. In Chapter 4, we explain the experiments conducted in this study and report the results. Chapter 5 describes the future directions of DICONS. We conclude in Chapter 6 with a brief summary.

2 RELATED WORK

2.1 Contextual Reference

In the field of natural language processing, coreference resolution is one of the tasks that deal with contextual reference. This task is the process of estimating the target of pronouns or demonstratives and complementing a zero pronoun, which is the omitted noun phrase. There are two basic methods to deal with coreference resolution: a rule-based method that extracts coreference relations based on the heuristic rules (Lee et al., 2011), and a method that uses machine learning (Clark et al., 2019; Joshi et al., 2019; Lee et al., 2018; Joshi et al., 2020).

The major problem with previous methods is that they do not consider situational reference. That is, they cannot handle a case in which the object that a pronoun refers to exists in visual information. Furthermore, whole sentences are required to resolve coreferences, and the sequentiality of dialogue is not considered. As stated earlier, the term "sequentiality of dialogue" refers to how the context gradually develops according to the utterance of each speaker. In order for the system to understand the utterance and generate a reply in a multi-turn conversation, it needs to dynamically consider possible reference candidates from the history of dialogue. Estimating the reference candidate will inevitably involve uncertainty. Furthermore, a person might make an additional utterance after the estimation has started, which is also an important hint for estimating the reference candidates, so the system has to continuously look back on the dialogue and reinterpret the reference candidate to reduce the uncertainty. In order to achieve meaningful communication between people and a system in the real world, it is necessary for the system to be able to consider the sequentiality of dialogue.

2.2 Situational Reference

A referring expression (RE) is any noun phrase, or surrogate for a noun phrase, whose function in discourse is to identify some individual object. In human conversations, humans interpret referring expressions using language, gesture, and context, fusing information from multiple modalities over time. Interpreting Multimodal Referring Expressions (Whitney et al., 2016) and Interactive Picking System (Hatori et al., 2018) deal with the RE problem, which is a task to identify objects in an image that correspond to ambiguous instructions.

Visual dialog (VisDial) (Das et al., 2017) is a task which requires a dialog agent to answer a series of questions grounded in an image. Attention-based approaches were primarily proposed to address these challenges, including Dual Attention Networks (Kang et al., 2019) and Light-weight Transformer for Many Inputs (Nguyen et al., 2020). VisDial dataset is a standard dataset for evaluating methods dealing with visual dialog. This dataset has two components: image and dialogue history about the image. A dialogue history is a set of successive question-answer pairs.

This paper focuses on more realistic dialogues than previous works. The setting in RE problem assumes that the referent must exist in the given image. In other words, the setting considers only situational references and excludes contextual references. VisDial dataset has also several disadvantages. First, the dialogue history consists of only questions about an image and answers for them, but realistic dialogue does not necessarily consist of such question-answering pairs. For example, one may give his/her thoughts about surrounding objects or opinions on others' utterance, so it is not natural to ask questions only. Second, all the questions in a dialogue history

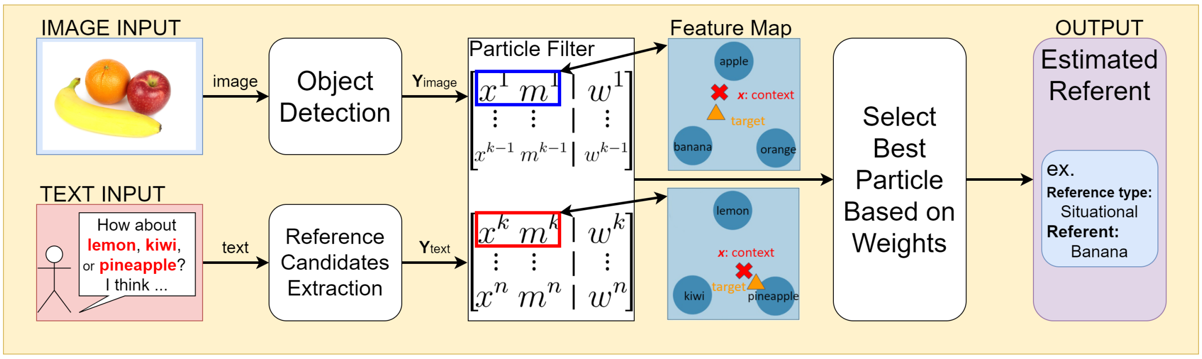

Figure 2: Flow chart of DICONS.

are about the things in a corresponding image. This means that no contextual references are taken into account. In a realistic situation, the situation where either contextual reference or situational reference occurs is limited. Since it is usually unknown which reference will occur, it is desirable that one system can handle both references.

Therefore, this paper deals with situations that include both contextual and situational references, which are closer to realistic situations. Since there was no dataset to deal with this setup, we prepared a few samples to evaluate our system on a trial basis.

2.3 SCAIN

SCAIN (Takimoto et al., 2020) is an algorithm that dynamically estimates the context and interpretations of words in a conversation. SCAIN achieves parallel context estimation while being able to obtain new information throughout the sequential input of a conversation text. As such, SCAIN can deal with the sequentiality of dialogue in dynamic resolution of coreference.

SCAIN evolved from FastSLAM (Montemerlo et al., 2002), an algorithm for a mobile robot that statistically resolves the interdependence between the robot's self-position and a map. SCAIN replaces self-position with a context and the map with a word interpretation space to apply FastSLAM to the interdependence between context and interpretation in a multi-turn conversation. The Kalman filter and a particle filter are the key mechanisms brought over from FastSLAM.

In SCAIN, the following processes are performed to sequentially interpret a dialogue and estimate the context. The correspondences between the SLAM and SCAIN variables are listed in Table 1.

SCAIN consists of 3 steps. First is update self-position, second is update landmark points, and third is resampling of particles. In step 1, SCAIN applies the interpreter's own utterance u to the context x in each particle. In step 2, SCAIN interprets the other's utterance z on the interpretation domain m and updates m. In step 3, SCAIN reflects the relationship between z and x. Then, SCAIN leaves valid interpretations by resampling the particles.

Table 1: Correspondence of random variables between conventional SLAM and SCAIN.

Random variable	Conventional SLAM	SCAIN
x	Self-position	Context
m	Environment map	Interpretation domain
u	Control	Own utterance
z	Observation	Other's utterance

3 DICONS

We propose Dynamic and Incremental Interpretation of Contextual and Situational References in Conversational Dialogues (DICONS) as a method for dynamically estimating context and word interpretation in a multi-turn conversation, considering both contextual information and visual information. With DICONS, it is possible to sequentially solve the MRI problem; it can determine whether the reference type in the dialogue is a contextual reference or a situational reference and estimate the referent of a demonstrative word.

At the timing when the demonstrative word appears in the conversation, DICONS starts to interpret the demonstrative word based on visual and contextual information. DICONS handles two types of particle: one for contextual reference and the other for situational reference, where both types of particle can handle conversational sentences sequentially. The difference between the two is that particles for contextual reference handle only utterance information whereas particles for situational reference handle both utterance information and images showing visual information of the space in which the conversation is occurring. The input image is used to obtain the candidates of a demonstrative word.

Candidates of a demonstrative word include two types: one that can be obtained from a conversation history $\mathbf{Y_{text}} = \{\mathbf{y_{1(text)}}, \mathbf{y_{2(text)}}, ...\}$ and one that can be obtained from an image $\mathbf{Y_{image}} = \{\mathbf{y_{1(image)}}, \mathbf{y_{2(image)}}, ...\}$. Here, each element of $\mathbf{Y_{text}}$ and $\mathbf{Y_{image}}$ is a distributed representation (Mikolov et al., 2013) of each interpretation candidate. The particles for contextual reference treat the words extracted from the past conversation history as the interpretation candidates. The candidates from situational reference are assumed to appear in the input image, and multiple labels extracted by object detection (Redmon et al., 2016) are treated as the candidates from the situational reference.

In a conversational sentence, the utterances after the appearance of the demonstrative word are input to each particle as the other's utterance. Following the steps 1–3 of SCAIN, the context and the distributed representations of the words, which are included in the interpretation domain, are updated. In detail, the update method of context is different from SCAIN. In step 1 of SCAIN, the number of particles is increased by randomly dividing the context of one particle into multiple directions, and it achieves an efficient context search in the distributed representation space. The update expression for the context $x_{t+1}^{k \cdot n+i}$ in the k-th particle at time t is represented, for i in the range from 1 to n of $\mathbf{Y}$, as

$$x_{t+1}^{k \cdot n+i} = (1-\lambda_u)x_t^k + \lambda_u v_{ut} + \sigma_i \ , \qquad (1)$$

where n is the number of interpretation candidates in $\mathbf{Y}$, $\mathbf{Y} = (y_1, \dots, y_i, \dots, y_n)$, $\mathbf{Y} = (\mathbf{Y_{text}} \text{ or } \mathbf{Y_{image}})$, λ_u is a hyper-parameter, σ_i is random Gaussian noise, and v_u is the weighted average of the distributed representation of the words composing the utterance u.

However, in DICONS, the number of particles is increased by dividing the context into the directions of the distributed representations of the candidates handled by the particles. This is represented as

$$x_{t+1}^{k \cdot n+i} = (1-\lambda_u)x_t^k + \lambda_u v_{ut} + \alpha y_i \ , \qquad (2)$$

where α is a hyper-parameter.

We do this so as to achieve a more efficient search in the distributed representation space than the random method can by assuming that the context is related to the candidates, as the candidates of the demonstrative word are limited in the task of estimating the referent. Then, as a result of the resampling in step 3 of SCAIN, the weight of each particle w^k is calculated based on the context and the interpretation domain, and the particle with the largest weight is selected as the optimum interpretation. The system can judge whether the contextual information or the visual information is dominant based on whether the best particle is derived from the context or the situation.

Next, the system decides the referent of the demonstrative word. This is done by comparing the cosine similarity between the target TGT of the best particle obtained by resampling and each candidate, and the candidate E with the maximum cosine similarity is selected as the referent. Here, TGT is the tracker of the possible interpretation candidates of a demonstrative word included in interpretation domain m. The above process can be represented by Eq. 3 below.

$$E = \underset{y_i \in \{\mathbf{Y_{text}} \text{ or } \mathbf{Y_{image}}\}}{\text{argmax}} \{\text{cosine\_similarity}(y_i, TGT)\} \qquad (3)$$

Note that $\text{cosine\_similarity}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}}{|\mathbf{x}|} \cdot \frac{\mathbf{y}}{|\mathbf{y}|}$, where $\mathbf{x}, \mathbf{y}$ are vectors.

4 CASE STUDY

4.1 Evaluation Method

To evaluate whether the proposed DICONS can solve the MRI problem in a sequential way, we conducted a case study of estimating the referent of demonstrative words. The purpose of the experiment is to consider whether it is possible to 1) determine the reference type of the demonstrative word and 2) estimate the referent.

In this experiment, two types of conversation examples with different difficulties were designed. Conversation example 1 is (Tables 2 and 3) an example in which the candidates for contextual reference and the candidates for situational reference are quite different such as fruits and birds. Conversation example 2 (Tables 6 and 7) is an example in which the candidates for contextual reference and the candidates for situational reference are similar such as fruits and fruits, so it is assumed to be more difficult to distinguish than conversation example 1.

Conversation examples are dialogues between two speakers: "A" and "B". Here, "A" plays a role of promoting the utterance, while "B" provides information useful for estimating the referent. In our case study, our system observes the utterances of "B" and estimates the referent. Interpretation candidates and the correct referent of a demonstrative word are shown in Tables 4 and 5 for conversation example 1 and Tables 8 and 9 for conversation example 2. The demon-

Table 2: Conversation example 1.1. The ground truth of the demonstrative word is as follows. Reference type: Contextual. Referent: Mango.

Speaker	Utterance content
A	“Bananas and pineapples are grown here. This is the first time I’ve actually seen any of them. Oh, mangoes are eaten by birds. I like mangoes.”
B	“I like them, too.” (∗timing to get interpretation candidates)
A	“Really?”
B (input_tx1a)	“They are sweet and delicious, aren’t they?”
A	“Yes, they are.”
B (input_tx2a)	“There are various desserts made from them. I like eating them as jelly.”

Table 3: Conversation example 1.2. The ground truth of the demonstrative word is as follows. Reference type: Situational. Referent: Crow.

Speaker	Utterance content
A	“Bananas, pineapples, and mangoes are grown here. This is the first time I’ve actually seen any of them.”
B	“This botanical garden also has many kinds of birds. There are birds everywhere.”
A	“I like that bird.” (∗timing to get interpretation candidates)
B (input_vis1a)	“Are you talking about the black one?”
A	“Yes.”
B (input_vis2a)	“Well, I think it’s cool in shape and color, but I don’t like it because it digs in the trash in the city.”

strative words in the conversation examples are interpreted following the algorithm in Sec. 3.

4.2 Conversation Example 1: Easy-to-Identify Examples

4.2.1 Evaluation Results

In each conversation example, the change in cosine similarity between the target and each interpretation candidate in the best particle and the change of reference type of the best particle are shown in Tables 4 and 5. Input_tx1a, tx2a in Table 4 and input_vis1a, vis2a in Table 5 correspond to the sentences shown in the Tables 2 and 3. The cosine similarity in Tables 4 and 5 refers to the initial position and the best particle calculated after each input. Here, regarding the initial position, since there is no difference in the weight between particles when there is no input, the confidence is the same for all interpretation candidates. The cosine similarity is regarded as equal to the confidence, and 0.5 is set as the initial position value. Note that 0.5 indicates the ambiguous state where it is unknown if each interpretation candidate is a referent or not. The best particle is one particle with the highest weight. Cosine similarities were calculated for either contextual or situational interpretation candidates because one particle considers only either ones. Therefore, there are some empty spaces (–) in Tables 4 and 5.

Table 4: Cosine similarity between each candidate and target of best particle (excluding initial position) and estimation result in conversation example 1.1. ∗ indicates the ground truth.

Candidates / Input	Initial position	input_tx1a	input_tx2a
Cosine similarity of contextual candidates			
mango∗	0.5000	0.9973	0.9722
banana	0.5000	0.7082	0.8478
pineapple	0.5000	0.8583	0.8439
Cosine similarity of situational candidates			
crow	0.5000	–	–
pigeon	0.5000	–	–
sparrow	0.5000	–	–
Estimation result			
Reference type	–	contextual	contextual
Estimation	–	mango	mango

Table 5: Cosine similarity between each candidate and target of best particle (excluding initial position) and estimation result in conversation example 1.2. ∗ indicates the ground truth.

Candidates / Input	Initial position	input_vis1a	input_vis2a
Cosine similarity of contextual candidates			
mango	0.5000	–	–
banana	0.5000	–	–
pineapple	0.5000	–	–
Cosine similarity of situational candidates			
crow∗	0.5000	0.9771	0.9256
pigeon	0.5000	0.4601	0.4334
sparrow	0.5000	0.4112	0.7026
Estimation result			
Reference type	–	situational	situational
Estimation	–	crow	crow

In addition, in each conversation example, in order to analyze whether the reference types are correctly distinguished, we visualized the ratio of the reference types in the top-30 particles, which is 10% of the total number of particles, and the distribution of the particle weights. The change in the ratio of the reference type (context or image) in the top-30 particles in descending order of weight is shown in Figs. 3 and 4. The change in the distribution of particle weights is

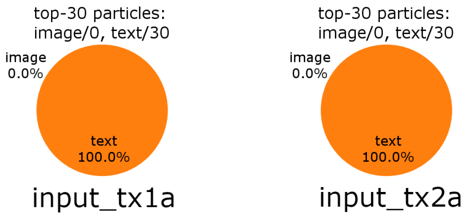

Figure 3: Change in ratio of reference type in top-30 particles in conversation example 1.1. No particles derive from the situational reference in the top-30 particles.

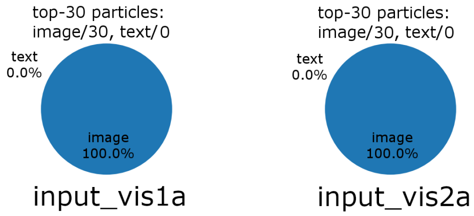

Figure 4: Change in ratio of reference type in top-30 particles in conversation example 1.2. No particles derive from the contextual reference in the top-30 particles.

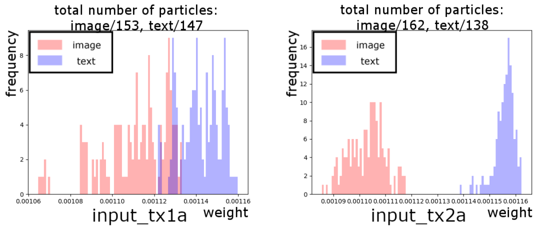

Figure 5: Change in distribution of particle weights in conversation example 1.1. The vertical line shows the frequency (that is, the number of particles per weight) and the horizontal one the weight of particles.

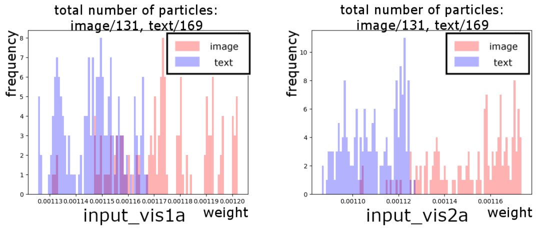

Figure 6: Change in distribution of particle weights in conversation example 1.2.

shown in Figs. 5 and 6.

4.2.2 Discussion

In conversation example 1.1, the column of input_tx2a in Table 4 has the highest cosine similarity. In conversation example 1.2, the column of input_vis2a in Table 5 has the highest cosine similarity. Therefore, the DICONS estimation of conversation example 1.1 is “mango”, which comes from the context, and in conversation example 1.2 it is “crow”, which comes from the input image. In both cases, the referent of the demonstrative word was correctly obtained.

From Figs. 3 and 4, there are only particles with the correct reference type in the top-30 particles. In both conversation examples, from Figs. 5 and 6, the weight distributions of the particles were clearly separated by the reference type in the progress of the dialogue. Therefore, our system can clearly distinguish the reference types.

4.3 Conversation Example 2: Difficult-to-Identify Examples

4.3.1 Evaluation Results

In each conversation example, the change in cosine similarity between the target and each interpretation candidate in the best particle and the change of reference type of the best particle are shown in Tables 8 and 9. Input_tx1b, tx2b in Table 8 and input_vis1b, vis2b in Table 9 correspond to the sentences shown in the Tables 6 and 7.

In addition, in each conversation example, the change in the ratio of the reference type (context or image) in the top-30 particles in descending order of weight is shown in Figs. 7 and 8. The change in the distribution of particle weights is shown in Figs. 9 and 10.

4.3.2 Discussion

The DICONS estimation of conversation example 2.1 is “lemon”, which comes from the context, and in conversation example 2.2 it is “banana”, which comes from the input image. In both cases, the referent of the demonstrative word was correctly obtained.

From Figs. 7 and 8, the correct reference type become dominant due to the sequential update. In the conversation example 2.1, from Fig. 9, the weight distribution of the situational reference particles is smaller and that of the contextual reference particles is larger in the progress of the dialogue from input_tx1b to input_tx2b. In the conversation example 2.2, from Fig. 10, the weight distribution of the contextual reference particles moves toward a small direction in the progress of the dialogue from input_vis1b to input_vis2b. Therefore, the estimation correctly changes even when focusing on all particles.

5 FUTURE WORK

This paper has some future works. For example, use of other visual information contained in the image added to the labels, evaluation on dataset or interac-

Table 6: Conversation example 2.1. The ground truth of the demonstrative word is as follows. Reference type: Contextual. Referent: Lemon.

Speaker	Utterance content
A	"I have a cold. Do you know any fruit that is good for colds?"
B	"How about kiwi or pineapple?"
A	"Well, anything else?"
B	"How about lemon?"
A	"Oh, it is certainly effective." (∗timing to get interpretation candidates)
B (input_tx1b)	"Yeah, it is rich in vitamin C and citric acid, so I think it is effective."
A	"OK, I must buy it."
B (input_tx2b)	"Yeah, but, It is sour and hard to eat, so I recommend eating it with other foods."

Table 7: Conversation example 2.2. The ground truth of the demonstrative word is as follows. Reference type: Situational. Referent: Banana.

Speaker	Utterance content
A	"I have a cold. Do you know any fruit that is good for colds?"
B	"How about lemon, kiwi, or pineapple?"
A	"Good. What do you think of that fruit?" (∗timing to get interpretation candidates)
B (input_vis1b)	"It has a lot of sugar and changes quickly to energy."
A	"So, do you think it is effective for colds?"
B (input_vis2b)	"Yes. Better with bread or cereal."

tion experiments with humans, and the well-designed method to obtain the contextual reference candidates.

6 CONCLUSIONS

The biggest problem with previous methods for demonstrative word interpretation is that they cannot deal with MRI problem. Our experimental results demonstrate that DICONS can handle the task with MRI problem in a multi-turn conversation.

ACKNOWLEDGEMENTS

This work was supported by JST CREST Grant Number JPMJCR19A1, Japan.

Table 8: Cosine similarity between each candidate and target of best particle (excluding initial position) and estimation result in conversation example 2.1. ∗ indicates the ground truth.

Candidates / Input	Initial position	input_tx1b	input_tx2b
Cosine similarity of contextual candidates			
lemon∗	0.5000	0.9931	0.9559
kiwi	0.5000	0.4078	0.6015
pineapple	0.5000	0.6717	0.6891
Cosine similarity of situational candidates			
banana	0.5000	–	–
apple	0.5000	–	–
orange	0.5000	–	–
Estimation result			
Reference type	–	contextual	contextual
Estimation	–	lemon	lemon

Table 9: Cosine similarity between each candidate and target of best particle (excluding initial position) and estimation result in conversation example 2.2. ∗ indicates the ground truth.

Candidates / Input	Initial position	input_vis1b	input_vis2b
Cosine similarity of contextual candidates			
lemon	0.5000	–	–
kiwi	0.5000	–	–
pineapple	0.5000	–	–
Cosine similarity of situational candidates			
banana∗	0.5000	0.9760	0.9385
apple	0.5000	0.5311	0.6099
orange	0.5000	0.5999	0.7408
Estimation result			
Reference type	–	situational	situational
Estimation	–	banana	banana

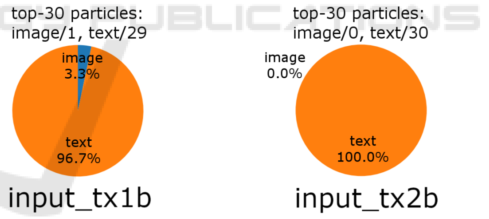

Figure 7: Change in ratio of reference type in top-30 particles in conversation example 2.1.

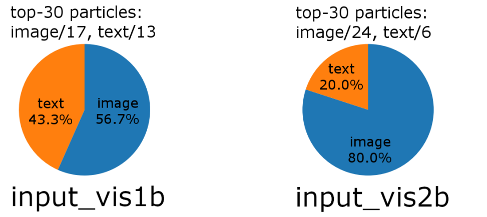

Figure 8: Change in ratio of reference type in top-30 particles in conversation example 2.2.

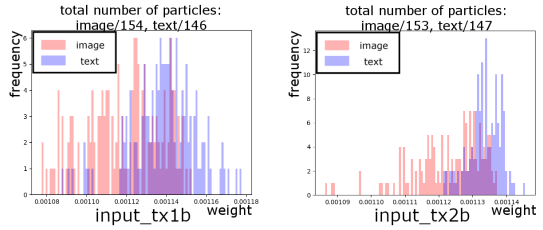

Figure 9: Change in distribution of particle weights in conversation example 2.1.

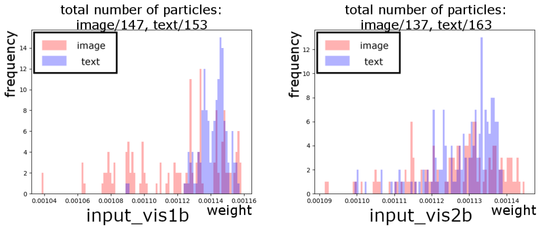

Figure 10: Change in distribution of particle weights in conversation example 2.2.

REFERENCES

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does BERT look at? an analysis of BERT's attention. *arXiv preprint arXiv:1906.04341*.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., Parikh, D., and Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W., and Tan, J. (2018). Interactively picking real-world objects with unconstrained spoken language instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Joshi, M., Levy, O., Weld, D. S., and Zettlemoyer, L. (2019). BERT for coreference resolution: Baselines and analysis. *arXiv preprint arXiv:1908.09091*.

Kang, G.-C., Lim, J., and Zhang, B.-T. (2019). Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2024–2033.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*, pages 28–34. Association for Computational Linguistics.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Montemerlo, M., Thrun, S., Koller, D., Wegbreit, B., et al. (2002). FastSLAM: A factored solution to the simultaneous localization and mapping problem. *AAAI/IAAI*, 593598.

Nguyen, V.-Q., Suganuma, M., and Okatani, T. (2020). Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *Proceedings of the 16th European Conference on Computer Vision*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Takimoto, Y., Fukuchi, Y., Matsumori, S., and Imai, M. (2020). SLAM-inspired simultaneous contextualization and interpreting for incremental conversation sentences. *arXiv preprint arXiv:2005.14662*.

Whitney, D., Eldon, M., Oberlin, J., and Tellex, S. (2016). Interpreting multimodal referring expressions in real time. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3331–3338. IEEE.