2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Operative Action Captioning for Estimating System Actions

Taiki Nakamura[1,2], Seiya Kawano[1], Akishige Yuguchi[1], Yasutomo Kawanishi[1], and Koichiro Yoshino[1]

Abstract— **Human-assistive systems, such as robots, need to correctly understand the surrounding situation based on observations and output the required support actions for humans. Language is one of the important channels to communicate with humans, and robots are required to have the ability to express their understanding and action-planning results. In this study, we propose a new task of operative action captioning that estimates and verbalizes the actions to be taken by the system in a human-assisting domain. We constructed a system that outputs a verbal description of a possible operative action that changes the current state to the given target state. We collected a dataset consisting of two images as observations, which express the current state and the state changed by actions and a caption that describes the actions that change the current state to the target state, by crowdsourcing in daily life situations. Then we constructed a system that estimates an operative action by a caption. Since the operative action's caption is expected to contain some state-changing actions, we use scene graph prediction as an auxiliary task because the events written in the scene graphs correspond to the state changes. Experimental results showed that our system successfully described the operative actions that should be conducted between the current and target states. The auxiliary tasks that predict the scene graphs improved the quality of the estimation results.**

I. Introduction

Recent advances in deep learning technology have fueled to research on situation understanding in which images, signals, and various other observations are understood through language [1], [2]. Systems can provide interpretations of data in a form comprehensible to humans by adding explanations to data through language for a variety of applications [3], [4]. There are various applications for such systems, one of which is systems that operate in human living spaces, such as life-support robots. Since these systems operate in a symbiotic space with humans, they must understand the situation in a form that can be interpreted by humans, such as natural language. Such language-based situational understanding has been discussed in a variety of situations, including human behavior analysis [5], describing robot behaviors [6], [7], and robot observations [8].

Correctly understanding and explaining a situation from observations is the first step in building a system that can work in human living spaces. However, systems are expected to provide cooperative assistance to human users. In other words, they must recognize both the current situation and the necessary actions (help) for solving it. For example, research in robotics has studied the problem of accurately identifying the expected robot action class given the current robot observation as input [9], [10]. Another study defined the problem of estimating the robot's action class that should be performed between the current and target states [11].

One of the most critical issues of these proposals is that there is a wide variety of life support systems that resemble those of robots, and the actions conducted for such support are also diverse. In other words, achieving a flexible understanding of a situation is challenging with only predefined robot action classes. Thus, in this study, we propose an operative action captioning task, which describes what action is to be done between the current to the target state by captioning for a better understanding of the surrounding human situation. One advantage of using captioning is that undefined action classes can also be generated by a language decoder. In this task, we assume that the current state (observation) and the target state are given by cameras. Then the system generates a caption that explains the expected actions to change the current state to the target state. Using this method, such human-assisting systems as robots can interact with users for mutual understanding.

Situation understanding by captioning has been actively studied in computer vision, which requires a certain amount of training data [12], [13], [14]. Therefore, we used crowdsourcing to construct a dataset of about 17,000 cases for operative action captioning. Existing research on captioning suggests the importance of using auxiliary information that can be recognized from images to improve the accuracy of captioning with limited training data [15], [16], [17]. We focus on scene graphs [18], which represent events on images as auxiliary information. The scene graph represents some events in the image, which are strongly related to the actions conducted in the images. A scene graph describes the situation in detail as a set of triplets: subject-relationship-object. By taking information from the scene graph for both the current and target states, we can acquire the differences or the conducted events between two states, which is critical information to acquire the action between them.

Our research contribution follows:

- We defined a new task, operative action captioning, toward building a robust robot action generation system.
- We built a new dataset for the defined task by extending

*A part of this work was supported by JSPS KAKENHI grant number 21H03519, 22H03654, and 22H04873.

[1]Guardian Robot Project, R-IH, RIKEN, 2-2-2, Hikaridai, Seika, Sohraku, Kyoto, 6190288, Japan {seiya.kawano, akishige.yuguchi, yasutomo.kawanishi, koichiro.yoshino} @riken.jp

[2]Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan supikiti@g.ecc.u-tokyo.ac.jp

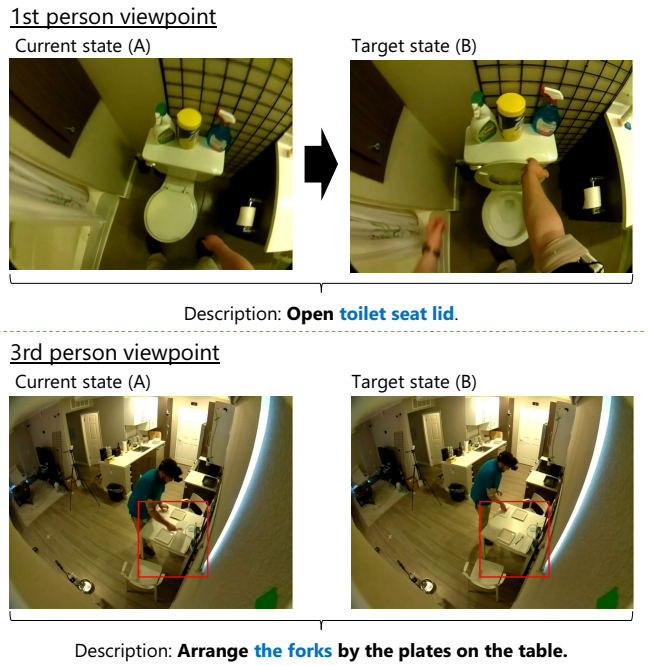

Fig. 1. Overview of the proposed task: operative action captioning. In the dataset, an assistant, who worked in the role of a human assisting a robot, has a head-mounted camera that is used to capture images. The first-person viewpoint indicates images from the assistant. The third-person viewpoint indicates images from fixed external cameras (such information can also be used by systems if external cameras are available). The red squares indicate areas that contain target objects for the operation.

the existing dataset, Home Action Genome Dataset.[1]

- We proposed a strong baseline based on change captioning and an auxiliary task of scene graph prediction.

II. Operative Action Captioning

Fig. 1 overviews the operative action captioning task. (A) is the current state observed by the life support system, such as a robot, and (B) is the target state when the support action is completed. In this case, "opening the toilet seat lid" is the expected support action, and the system has to distinguish it from both the current and target states. In this study, images (A) and (B) are used as input, and the action that changes the state (A) to (B) is estimated by generating explanatory sentences (caption).

In a general scenario for robots, the problem is defined as selecting an action class from pre-defined action classes, given these two inputs [11]. In other words, the problem is predicting an action class that can change the current state (A) to the target state (B). However, a wide variety of actions must be performed in daily life support. This diversity makes it difficult to estimate the support actions that are required for life support systems. We apply a generative approach based on captioning to estimate the necessary support actions from both pre-defined action classes and undefined actions.

Recognition of such operative actions has been researched in the field of possible action recognition from observations [19], [9], [20], [10]. Such studies define the problem by estimating the action class contained in images or videos. A more complicated event representation, such as scene graphs [18], is used to predict actions or events in observations [21], [22]. In this study, we introduce captioning using natural language as a more flexible estimation of operative actions. The use of scene graphs also generates accurate and clear captions [23]. Our study is inspired by such works and uses scene graph prediction as an auxiliary task to improve the generated caption. By using scene graph prediction as an auxiliary task, the system can benefit from the information obtained from the scene graph without having to perform scene graph prediction in advance during the testing phase.

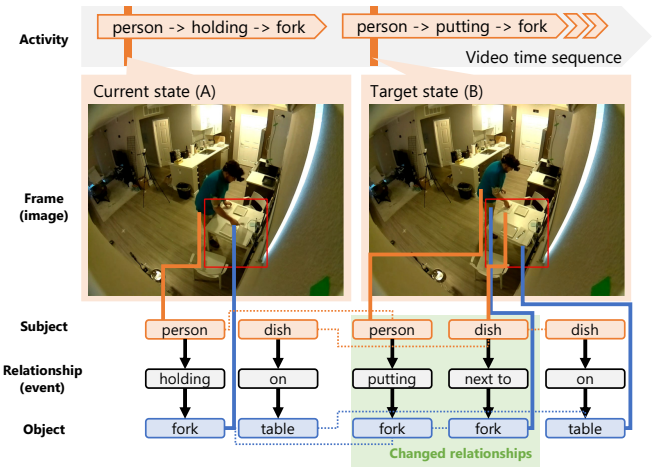

Fig. 2. Annotation example of Home Action Genome Dataset

Captioning differences is another related research direction that focuses on the different descriptions of two images [24], [25]. In this study, our task captions the actions that should be conducted between two states (images). The difference information is useful to predict the operative actions; thus, we use a captioning system of difference as our baseline.

Raw-level robot action estimation from two such observations is another research avenue [26]. A robot's physicality strongly constrains such an approach. Our captioning approach can express the expected action by language, even if the target robot cannot conduct the generated action. This point is critical for robots working in human living space.

III. Data Collection

We collected texts that describe operative actions for pairs of the current and target states (images) for the estimation based on captioning, because our system requires training captions that focus on the conducted operative actions between two states. We extended the Home Action Genome Dataset [22] to achieve operative action captioning for developing a robot at home. In this section, we describe the Home Action Genome Dataset, our data collection method using crowdsourcing, and our collection results.

A. Home Action Genome Dataset

The Home Action Genome Dataset [22] consists of videos of various human actions in the home with their annota-

[1]Our data will be available at here after the conference: https://github.com/riken-grp/operative_action_captioning

tions (Fig. 2): subjects, objects, and relationships. Relationships contain relations, events, and actions. The videos are recorded by head-mount cameras for a first-person viewpoint and fixed-point cameras for a third-person viewpoint. The third-person videos have annotations in the form of scene graphs that describe events or actions with object names, subject names, and rectangles. scene graphs indicate "subject-relationship-object" connections. "Activities" are defined as events related to operative actions. These annotations are given with a time stamp in the video's time series. Since both the third- and first-person videos are recorded in all the sessions and are time-synchronized, these labels can be used for both videos even though we cannot use a rectangle of the object area in the first-person video. We used both the third- and first-person videos for our task, because human assisting systems such as robots can use both viewpoints.

Some specific annotation examples are shown in Fig. 2. Here scene graphs are annotated to corresponding events and relationships to the action: putting the forks on the table by the plate. In the actual annotation, although such relationships as "person-in_front_of-table" are comprehensively annotated, the example in Fig. 2 only shows the graphs related to the target operative action.

B. Caption Annotation by Crowdsourcing

To achieve operative action captioning, we used frames in the Home Action Genome Dataset videos. The current state is defined as an observation just before an actual operative action, and the target state is defined as an observation just after it. The current and target states are defined by extracting image frames before and after the point in a time sequence when the relationship changed that corresponds to the activity in the scene graph. For example, in the example in Fig. 2, we extracted the frames before and after the "person-holding-fork" scene graph changed to "person-putting-fork." We automatically extracted approximately 69,000 pairs of candidate frames using scene graph annotation from the Home Action Genome Dataset from both first- and third-person viewpoint videos. For each of these pairs, we added by crowdsourcing a natural language description of what kind of operative action was performed between the paired images. We presented both images and the target object name's label extracted from the scene graph annotation to the crowd workers and asked them to explain what kind of action was performed by the target object between the images. The crowd workers received the following instructions.

> What did the worker in the images do to change the state in the first image to the one in the second image? If you can explain using the "OBJECT", check the "I can explain" box and describe it using the object name. If you cannot explain, check the "I cannot explain" box and describe why.

OBJECT is the name of an object that is related to the target activity. For example, in Fig. 1, we used "toilet" as the OBJECT for the upper example and "fork" as the OBJECT for the lower example. Fig. 1 also indicates examples of collected captions as "description."

Finally, we gathered 16864 pairs with operative action captions and split them into 14335/843/1686 as train/valid/test. We checked the quality of the captions and the reasons for being unable to explain and recollected if the checking rejected the sample. The image pairs that were finally judged as unable to be annotated included the following:

- The target object is too far away to be identified by the third-person viewpoint camera.
- The position or state of the target object did not appear to change, between the current and target images.
- Since the target object is obstructed by a person or other objects, distinguishing it from the image is difficult.
- The first-person viewpoint image is blurred.
- The target object is not shown in the current image, in the target image, or in either image.

IV. Operative Action Estimation Based on a Natural Language Generation Model

Using the collected data, we constructed a model to estimate the operative actions performed (or to be performed) from the images of the current and target states. We used the dual dynamic attention (DUDA) model as our baseline scheme and improved it by adding a scene graph prediction module as an auxiliary task. In this section, we describe the outline of the DUDA model, the scene graph prediction used as the auxiliary task, and the training setup. The overall model overview is shown in Fig. 3.

A. Dual Dynamic Attention (DUDA) Model

The DUDA model [24], which focuses on the change between two images, was proposed to explain the change itself. Thus, it has an image-difference detection mechanism. We use this model because the part of the image change extracted by difference detection corresponds to an object that changes due to the operative action, which is our task's main focus. In the model, images before and after the operative action are converted into feature vectors by an image encoder using ResNet [27] to perform difference detection. The extracted feature matrices of images $X_{(\mathsf{A})}$ and $X_{(\mathsf{B})}$ calculate the difference matrix as $X_{\mathsf{diff}} = X_{(\mathsf{A})} - X_{(\mathsf{B})}$. The difference matrix is concatenated with the original matrices as $X'_{(\mathsf{A})}$ and $X'_{(\mathsf{B})}$ to compute spatial attentions $a_{(\mathsf{A})}$ and $a_{(\mathsf{B})}$ [28]. The elemental unit multiplication between the attention weights and feature matrices $X_{(\mathsf{A})}$, $X_{(\mathsf{B})}$, and X_{diff} are used as resultant feature vectors $l_{(\mathsf{A})}$, $l_{(\mathsf{B})}$, and l_{diff}, which focus on the changes between images. The change is also related to the operative action. The weighted feature vectors $l_{(\mathsf{A})}$, $l_{(\mathsf{B})}$, and l_{diff} are fed to the captioning network that has dynamic attention to generate captions. Here, our target caption indicates the conducted operative action between two images. The softmax cross-entropy loss L_{cap} for the reference caption is used as the loss function for network training. The original DUDA model does not contain the "auxiliary task" part in Fig. 3.

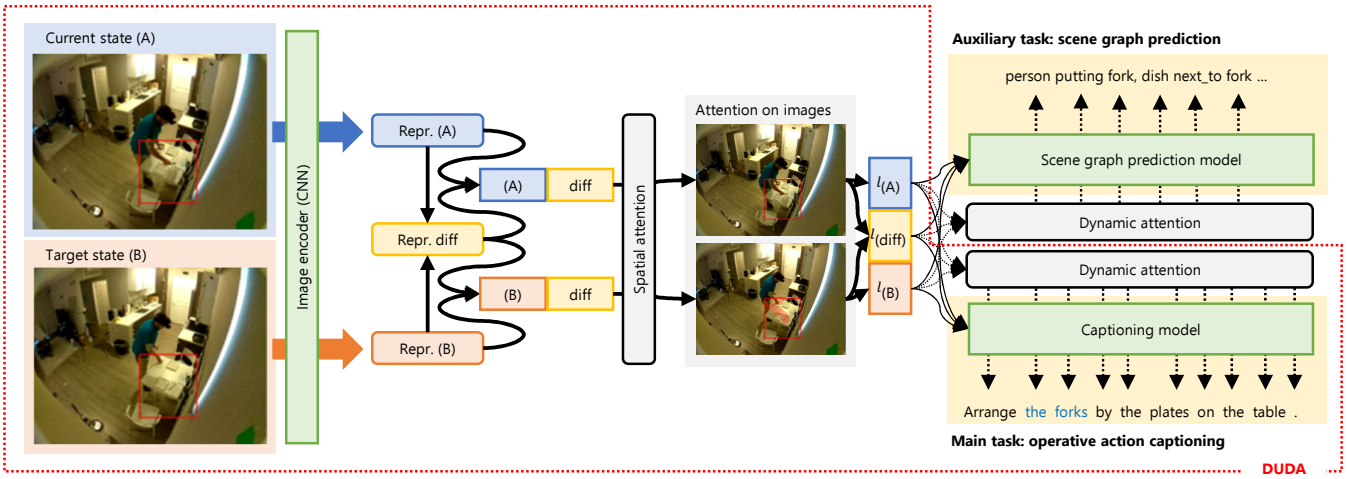

Fig. 3. Dual dynamic attention (DUDA) model and its extension in the proposed system

B. Auxiliary Task: scene graph Prediction

The actions performed between the current and target images are given as captions, which contain various expressions. Some captions do not explicitly contain the names of the target actions or the target objects. The auxiliary tasks that predict the scene graphs may facilitate the preservation of such information to improve the caption quality.

We accomplished this idea by adding a model for the auxiliary tasks that predicts scene graphs with the baseline DUDA model. As shown in Fig. 3, we constructed a model to predict the triples in scene graphs given the same image features to the captioning model: $l_{(\mathrm{A})}$, $l_{(\mathrm{B})}$, and l_{diff}. Although many scene graphs are obtained from the images, we built a sequential prediction model that repeatedly outputs "subject-relationship-object" by the network as shown in Fig 3. Since this auxiliary task performs sequential prediction, we defined the network that uses the softmax cross-entropy loss as in the main task. Let L_{sgr} denote the loss function of the auxiliary task.

C. Loss Function and Experimental Setup

We used the DUDA model without a scene graph as the baseline to capture the operative actions performed between both images. We used the original implementation of the DUDA model[2]. The initial learning rate was 0.01, which was multiplied by 0.1 every 20 epochs. The loss function is defined as,

$$L_\theta = L_{cap} + \lambda_{L_{1,cap}} L_1 - \lambda_{ent} L_{ent,cap}. \quad (1)$$

Here $L_{1,cap}$ and $L_{ent,cap}$ are regularizations. λ_{L_1} and λ_{ent} are the hyperparameters of the weights for each regularization. We use the same hyperparameters from the original DUDA implementation and denote this setting as Baseline.

When we trained the proposed model that has an auxiliary task to predict scene graphs, we extended the original DUDA implementation (Eq. (1)):

$$\begin{aligned} L(\theta) =& \alpha(L_{cap} + \lambda_{L_1} L_{1,cap} - \lambda_{ent} L_{ent,cap}) \\ &+ (1-\alpha)(L_{sgr} + \lambda_{L_1} L_{1,sgr} - \lambda_{ent} L_{ent,sgr}). \quad (2) \end{aligned}$$

Here α is a weight that integrates the weights for operative action captioning and scene graph prediction. $L_{1,sgr}$ and $L_{ent,sgr}$ are regularizations for the auxiliary task. We used the same hyperparameters for the baseline for λ_{L_1} and λ_{ent}.

We implemented two integration methods for the main and auxiliary tasks. The first is a linear interpolation that uses a fixed α. We set $\alpha = 0.9$ based on a trial on the validation dataset. We call this setting lin. (0.9). Another method is a fluctuating update, which alternatively uses two αs every ten epochs. We tried two αs patterns, $[0.0, 1.0]$ and $[0.1, 0.9]$, and called these settings alt. (1.0) and alt. (0.9). In the alternative method, we trained the scene graph prediction task and then the captioning task.

In addition, we utilized two scene graph sets to investigate the best usage. Since there are several scene graphs in the current state (A) and the target state (B), we determined two setups: all and diff. The all method predicts any scene graphs contained in both (A) and (B). The diff method predicts only the discrepancies between scene graphs (A) and (B). The former corresponds to using features from both images, and the latter corresponds to using the changes in the images.

V. Experiments

To evaluate each model described in the experimental setup, we conducted an automatic evaluation based on a comparison with the reference and a human evaluation in which the generated results were evaluated manually. The evaluation criteria and results are described below.

A. Evaluation Criteria

The automatic evaluation criteria are based on a systematic comparison with the annotated caption reference of the test set. We used BLEU-1_4 [29], which is based on the n-gram match rate, ROUGE-L [30], which is based on the maximum

[2] https://github.com/Seth-Park/RobustChangeCaptioning

TABLE I

AUTOMATIC EVALUATION RESULTS

Model	Condition		BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L	CIDEr
Baseline	-		0.389	0.238	0.151	0.0998	0.375	0.871
+SG	alt. (1.0)	all	0.350	0.194	0.113	0.0686	0.330	0567
		diff	0.359	0.205	0.126	0.0815	0.339	0.649
	alt. (0.9)	all	0.396	0.244	0.156	0.105	0.383	0.921
		diff	0.392	0.245	0.158	0.107	0.389	0.913
	lin. (0.9)	all	**0.405**	**0.260**	**0.167**	**0.114**	**0.392**	**1.001**
		diff	0.396	0.246	0.160	0.109	0.387	0.971

TABLE II

CONTENT WORD ACCURACY. NOTE THAT MINOR DIFFERENCES IN TERMS ARE IGNORED IN THIS EVALUATION.

Model	Condition		Noun P	Noun R	Noun F	Verb P	Verb R	Verb F	Verb-independent P	Verb-independent R	Verb-independent F
Baseline	-		0.378	0.385	0.381	0.156	0.169	0.162	0.122	0.131	0.126
+SG	alt. (0.9)	all	0.393	0.406	0.399	0.168	0.186	0.177	0.139	0.153	0.146
		diff	0.387	0.406	0.396	0.164	0.173	0.168	0.129	0.138	0.133
	lin. (0.9)	all	**0.398**	**0.414**	**0.406**	**0.176**	**0.190**	**0.183**	**0.144**	**0.158**	**0.151**
		diff	0.387	0.413	0.400	0.165	0.180	0.172	0.140	0.153	0.146

TABLE III

SCENE GRAPH PREDICTION PERFORMANCE IN PROPOSED MODELS

Cond.		Entity P	Entity R	Entity F	Triplet P	Triplet R	Triplet F
alt. (0.9)	all	0.679	0.694	0.673	0.567	0.564	0.542
	diff	0.664	0.677	0.654	0.508	0.424	0.429
lin. (0.9)	all	0.726	0.726	0.710	0.602	0.593	0.572
	diff	0.700	0.720	0.694	0.542	0.486	0.478

match length, and CIDEr [31], which is based on weighted-term matching.

We expect to add the information from the scene graph prediction to the captioning results of the motion behavior estimation; thus, we focused on nouns and verbs. We calculated precision (P), recall (R), and harmonic mean (F) of "nouns," "verbs," and "independent verbs" extracted from the references.

We also performed human evaluations by two human annotators because the correlations are limited between the automatic evaluation criteria and the human evaluation results. In the human evaluation, one human evaluator was given state images (images (A) and (B)) and generated captions that described their operative actions. The evaluator rated the caption's naturalness and informativeness [32] on a five-point scale. Another annotator checked and revised the first evaluator's result to improve the consistency. This process was done blindly; the evaluators did not know the method names. Because human evaluation is expensive, we evaluated 200 pairs randomly from the test set for the three systems with the best scores in the automatic evaluation: baseline, all, and diff for lin. (0.9).

B. Automatic Evaluation Results

Tables I and II show the automatic evaluation results. +SG indicate models using scene graph prediction as the auxiliary task. Both tables show a primary trend, where using scene graph predictions as auxiliary tasks improved each score compared to the baseline. The result suggests that linear interpolation with a small weight on the auxiliary task outperforms the alternative training.

Based on Table II, our proposed method successfully generated content words, including nouns and verbs, more often than the baseline. This is because the scene graph contents represented by "subject-relationship-object" facilitates the preservation of such information in the network.

In addition, scene graph prediction results, results of the auxiliary task, in proposed models were evaluated by their entity matching and triplet matching as shown in Table III. In this result, all in lin. (linear interpolation) condition achieved the best scores, which follow the general trend in automatic evaluation. We cannot directly compare all and diff conditions because they have different prediction targets.

C. Human Evaluation Results

Table IV shows the human evaluation results, which indicate that using scene graph prediction as the auxiliary task improved the informativeness and contributed significantly to the naturalness of either of our proposed methods (+SG with all or diff). This is probably because the auxiliary task suppressed the over-generation of content words that are not contained in the scene graphs. It might also suppress repetition, which is a typical problem of sentence generation, as shown in the case study analysis.

Fig. 4 shows the score distribution of each criterion. The naturalness was improved by focusing on the differences in scene graphs. Few differences exist between all and diff in informativeness, although they outperformed the Baseline. If we look at each case, the proposed method improved the informativeness in many cases, although sometimes the proposed method decreased it. Naturalness was improved by the proposed method in many cases.

TABLE IV

HUMAN EVALUATION RESULTS

Model	Condition		Natur.	Infor.
Baseline	-		3.89	3.05
+SG	lin. (0.9)	all	4.42	3.45
		diff	**4.53**	**3.48**

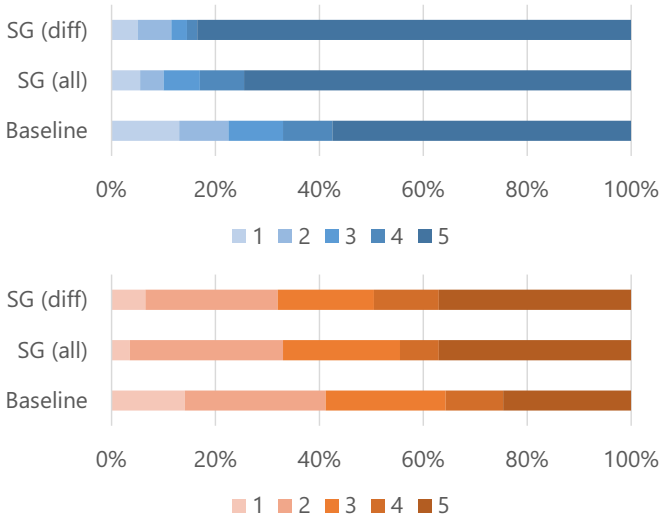

Fig. 4. Distribution of naturalness (upper) and informativeness (bottom) scores of each method

Fig. 5. Case study indicating generated samples

D. Case Study

As a case study, Fig. 5 shows four examples of the current state (A), the target state (B), and the captions from three methods: Baseline, +SG (all), and +SG (diff). In the first example, the baseline describes a meaningless and redundant action of "move the clothes in the basket to the basket," although the proposed methods with scene graph prediction as an auxiliary task generated actual actions. All, which looks at the entire scene graph, generated action "wash," and diff, which looks at the differences in the scene graphs, generated a more detailed action: "take out." Although both are correct, the former described the overall action in which the target action was included; the latter, which focused on the differences, described a specific action included in the idea of washing.

In the second example, the baseline generated an action that rarely occurs in reality, "put dishes with vegetables in bowls," which was improved in the proposed methods. However, when we focused on the differences (diff), the proposed method generated an opposite action "remove" instead of the actual action "place". Since opposite actions tend to be placed near each other in the embedding space, we must consider how to deal with such cases in future work.

In the third example, the baseline caused a repetition problem, which was suppressed by the proposed methods. Focusing on the difference might generate an object name "clothesline."

In the fourth example, both proposed methods successfully generated actions corresponding to using a dishwasher; however, they explained with different granularity. In particular, diff explained the "pull out the dishwasher basket" action, which is included as part of the overall "put the cup in the dishwasher" action. When we apply these methods to a human-assisting scenario at home, we must discuss the controllability of the granularity of the captions.

In the fifth example, only the baseline successfully captioned the action of "make coffee." This is probably because the proposed models tried to use objects in the scene graph and failed.

VI. Conclusion

We constructed a framework to verbalize the required operative actions given both the current and target (ideal) states by captioning networks to estimate the operative action of such human-assistive systems as robots. We constructed a dataset by crowdsourcing that consists of triplets of a current state, a target state, and a caption that describes the operative action to change the current state to the target state. We proposed a captioning model that uses scene graph prediction as an auxiliary task by focusing on object names and the relationships represented in scene graphs. We investigated the effect of our proposed method through both automatic and human evaluations, especially on human evaluation results on naturalness and informativeness. Our future work will include the proposed system in our human-assisting robot at home [8].

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[2] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4651–4659.

[3] L. Dou, G. Qin, J. Wang, J.-G. Yao, and C.-Y. Lin, "Data2text studio: Automated text generation from structured data," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, 2018, pp. 13–18.

[4] T. Ishigaki, G. Topić, Y. Hamazono, H. Noji, I. Kobayashi, Y. Miyao, and H. Takamura, "Generating racing game commentary from vision, language, and structured data," in *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*, 2021, pp. 103–113.

[5] W. Takano and Y. Nakamura, "Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions," *The International Journal of Robotics Research*, vol. 34, no. 10, pp. 1314–1328, 2015.

[6] T. Yamada, H. Matsunaga, and T. Ogata, "Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3441–3448, 2018.

[7] K. Yoshino, K. Wakimoto, Y. Nishimura, and S. Nakamura, "Caption generation of robot behaviors based on unsupervised learning of action segments," *Conversational Dialogue Systems for the Next Decade*, p. 227, 2020.

[8] A. Yuguchi, S. Kawano, K. Yoshino, C. T. Ishi, Y. Kawanishi, Y. Nakamura, T. Minato, Y. Saito, and M. Minoh, "Butsukusa: A conversational mobile robot describing its own observations and internal states," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2022, pp. 1114–1118.

[9] N. Soans, E. Asali, Y. Hong, and P. Doshi, "Sa-net: Robust state-action recognition for learning from observations," in *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2153–2159.

[10] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[11] R. Chatila, E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, *et al.*, "Toward self-aware robots," *Frontiers in Robotics and AI*, vol. 5, no. 88, 2018, DOI:10.3389/frobt.2018.00088.

[12] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," 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)*, 2019, pp. 2012–2023.

[13] R. C. Luo, Y.-T. Hsu, Y.-C. Wen, and H.-J. Ye, "Visual image caption generation for service robotics and industrial applications," in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, 2019, pp. 827–832.

[14] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata, and H. Kataoka, "Indoor scene change captioning based on multimodality data," *Sensors*, vol. 20, no. 17, p. 4761, 2020.

[15] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5630–5639.

[16] T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating copying mechanism in image captioning for learning novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6580–6588.

[17] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing novel objects in image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 497–12 506.

[18] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1261–1270.

[19] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6450–6459.

[20] X. Hong, Y. Lan, L. Pang, J. Guo, and X. Cheng, "Transformation driven visual reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6903–6912.

[21] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 236–10 247.

[22] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J. C. Niebles, "Home action genome: Cooperative compositional action understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 184–11 193.

[23] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9962–9971.

[24] D. H. Park, T. Darrell, and A. Rohrbach, "Robust change captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 4624–4633.

[25] Y. Qiu, Y. Satoh, R. Suzuki, K. Iwata, and H. Kataoka, "3d-aware scene change captioning from multiview images," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4743–4750, 2020.

[26] H. Kim, A. Zala, G. Burri, and M. Bansal, "Fixmypose: Pose correctional captioning and retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 13 161–13 170.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4942–4950.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.

[30] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[31] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 4566–4575.

[32] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1711–1721.