Modeling Collaborative Dialogue in Minecraft with Action-Utterance Model

Takuma Ichikawa and **Ryuichiro Higashinaka**
Graduate School of Informatics, Nagoya University, Japan
{ichikawa.takuma.w0@s.mail,higashinaka@i}.nagoya-u.ac.jp

Abstract

With the advancement of dialogue systems propelled by neural-based methods, researchers have been working on developing dialogue systems that can collaborate with humans to complete tasks in the real world and virtual environments. In such collaborative work, the system needs to either perform an action or make an utterance appropriate for the context. However, previous literature has treated action and utterance generation separately. In this study, with the aim of enabling the system to autonomously determine whether to act or utter, we create a model that can handle both action and utterance generation in a unified model. We conducted experiments on a dataset related to collaborative work in Minecraft and show that the proposed model can autonomously determine whether to act or utter and generate better actions and utterances than the baselines.

1 Introduction

With the advancement of dialogue systems by neural-based methods (Bang et al., 2023; Shuster et al., 2022), towards more advanced dialogue systems, researchers have been working on developing dialogue systems that can collaborate with humans to complete tasks (Meena et al., 2013; He et al., 2017). Many studies have focused on collaborative work in virtual environments, such as Minecraft (Narayan-Chen et al., 2019; Ogawa et al., 2020; Bara et al., 2021), and competitions such as the Interactive Grounded Language Understanding (IGLU) challenge[1] have been organized. In such collaborative work, systems need to handle not only dialogue but also actions in their environment. However, studies in previous literature treat action and utterance generation as separate tasks (Narayan-Chen et al., 2019; Jayannavar et al., 2020; Mohanty et al., 2023), making systems incapable of executing both, which is required in realistic settings.

In this study, with the aim of enabling a system to autonomously determine whether to act or utter and execute on the basis of context, we create a unified model, the Action-Utterance Model, that can handle both action and utterance generation. Specifically, the model is trained simultaneously on three tasks: action type classification, action generation, and utterance generation.

We conducted experiments using the Collaborative Garden Task Corpus (Ichikawa and Higashinaka, 2022), which is a dataset related to collaborative work in Minecraft, and the results showed that the proposed model can autonomously determine whether to act or utter and generate better actions and utterances than the baselines. Furthermore, we analyzed the inference results and revealed the difficulty of generating actions unrelated to last actions.

2 Related Work

Studies have been emerging on performing complex collaborative work involving both actions and utterances in virtual worlds such as Minecraft (Kim et al., 2019; Ichikawa and Higashinaka, 2022) with some implemented systems.

For example, Gray et al. (2019) constructed a system that creates simple structures on the basis of user instructions through text chat. Narayan-Chen et al. (2019) and Jayannavar et al. (2020) modelled an instructor and builder for the Collaborative Building Task (Narayan-Chen et al., 2019), which involves two interlocutors working together to create a target structure. Recent research has focused on the IGLU task, which is based on the Collaborative Building Task (Kiseleva et al., 2022; Mohanty et al., 2023; Shi et al., 2023; Mehta et al., 2023). However, while there have been efforts to classify whether to act or utter, these tasks are treated as

[1] https://www.iglu-contest.net/

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 75–81
November 1–4, 2023. ©2023 Association for Computational Linguistics

ID	S	Action or Utterance
1	A	なにか作りたいものありますか？ *(Do you want to make something?)*
2	B	藤？みたいな屋根みたいなのつくってみたいです *(I want to make a roof like a wisteria trellis.)*
	B	{(place, oak_fence, 4, -1, 4), (place, oak_fence, 4, -1, 5)}
3	A	いいですね！ *(Sounds good!)*
	B	{(place, oak_leaves, 4, -1, 6)}
4	A	真ん中にどーんと作ってみてください！ *(Try making it in the middle!)*
5	B	道を真ん中に作ってみます *(I will make a path down the middle.)*
	B	{(place, oak_leaves, 3, -1, 6), {(place, oak_leaves, 2, -1, 6), {(place, oak_leaves, 1, -1, 6), {(place, oak_leaves, 0, -1, 6), …}

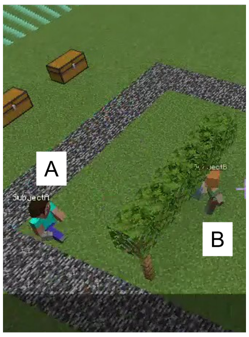


Figure 1: Dialogue in Collaborative Garden Task Corpus. ID represents utterance number, and S represents interlocutor. Utterances were originally in Japanese and have been translated into English by authors. Shaded rows indicate actions. Figure on right shows situation immediately after last action.

independent. Since actions and utterances are interrelated at the intent level and should not be treated separately, in this study, we focus on a model that can handle both action and utterance generation.

In fields outside of collaborative work, there are studies that aim to develop systems that can handle both actions and utterances. For example, Chen et al. (2021) constructed a task-oriented dialogue dataset that incorporates both actions, such as search and purchase, and utterances. Reed et al. (2022) proposed a model called Gato, which utilizes a single Transformer architecture to perform various tasks, including text generation tasks, such as utterance generation and caption generation, as well as action generation tasks. However, these studies do not use a model that can handle both actions and utterances. In this paper, we investigate the effectiveness of a unified model in collaborative work tasks.

3 Dataset and Task

3.1 Collaborative Garden Task Corpus

In this study, we adopt the Collaborative Garden Task Corpus constructed by Ichikawa and Higashinaka (2022) as a collaborative work dataset. Figure 1 shows an example of a dialogue included in the corpus. In the Collaborative Garden Task, two interlocutors interact via text chat while manipulating blocks in order to cooperatively create a beautiful and unique garden in Minecraft (here, beauty and uniqueness are based on the subjective evaluation of the interlocutors). In the dataset, the interlocutors can freely use 17 different types of blocks within a 10 × 10 × 4 area; they need to decide on the design of the garden through dialogue with their partner. Since the activity combines actions and utterances, we determined it to be a suitable dataset for evaluation. The Collaborative Garden Task Corpus contains 1,092 dialogues, each of which records in-game information such as utterances, block manipulations, and avatar movements. The language used is Japanese, with a total of 31,416 utterances, an average word count of 13.9 per utterance, and a total of 657,693 block manipulations.

3.2 Next Action-Utterance Generation Task

In this study, we address the Next Action-Utterance Generation Task, which aims to predict the next action or utterance to be performed. In this task, the goal is to predict the interlocutor's next actions (which may include making utterances), denoted as a_t, given the dialogue and state history H_t, the world state W_t, and the avatar's position (x_t, y_t, z_t) and orientation $(yaw_t, pitch_t)$ at turn t.

An action a_t is composed of one of four action types [utterance (UTT), block manipulation (BLOCK), SKIP, FINISH], along with its subsidiary information. In the case of UTT, we additionally predict utterance u_t. In the case of BLOCK, we additionally predict the set of block operations $b_t = \{(block\_action, [block\_name], x, y, z), ...\}$; $block\_action$ represents whether to place or break and (x, y, z) represents the block coordinates. In the case of placement, the block type $block\_name$ is also to be output. SKIP represents the interlocutors' non-operation at turn t. SKIP is introduced to model complex mixed-initiative interactions into a simple turn-by-turn dialogue. FINISH represents the end of the dialogue; the dialogue ends when one of the interlocutors outputs FINISH.

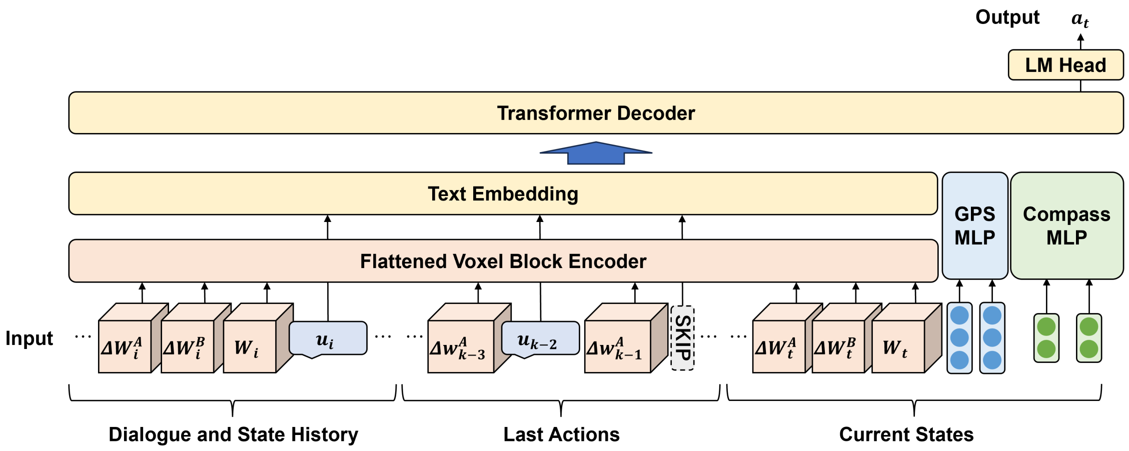


Figure 2: Overall architecture of proposed model. Model takes as input dialogue and state history $H_t = \{(\Delta W_{i_0}, W_{i_0}, u_{i_0}), ..., (\Delta W_i, W_i, u_i), ...\}$, actions (which may include making utterances) performed in last N turns, change in world state between previous utterance and current time, ΔW_t, current world state W_t, and avatar's positions and orientations. Model outputs next action type along with its content.

4 Model

4.1 Model Architecture

Figure 2 shows the overall architecture of the proposed model, the Action-Utterance Model. We use a pretrained Transformer decoder model as the underlying Large Language Model (LLM). Additionally, to embed non-verbal information, such as the world state and the avatar's position and orientation, into the same dimension as text, various encoders are prepared, including the Flattened Voxel Block Encoder, GPS multi-layer perceptron (MLP), and Compass MLP. These encoders were inspired by the implementation in MineDojo (Fan et al., 2022).

The Flattened Voxel Block Encoder consists of an embedding layer and a 3-layer MLP. It converts voxel data representing the world state into a vector equivalent to one token. The GPS MLP and Compass MLP consist of 2-layer MLPs, each transforming the avatar's positional and orientation information into a vector equivalent to one token. Each MLP is composed of linear and ReLU layers. Non-verbal information such as the world state embedded in the same vector space as the text is concatenated with the text embedding and input into the decoder. The LM Head receives the information processed by the decoder and outputs the next action.

4.2 Model Input

The model receives input at turn t, which includes the dialogue and state history, H_t, actions taken in the most recent N turns (we use $N = 10$ in this paper), the change in the world state between the previous utterance and the current time, ΔW_t, the current world state W_t, and the avatar's position (x_t, y_t, z_t) and orientation $(yaw_t, pitch_t)$.

The dialogue and state history consist of a set of tuples, ΔW_i, W_i, and the utterance u_i, formulated as follows.

$$H_t = \{(\Delta W_{i_0}, W_{i_0}, u_{i_0}), ..., (\Delta W_i, W_i, u_i), ...\} \quad (1)$$

ΔW_i represents changes in the world state between the previous utterance and the current utterance, while W_i represents the world state at the time of the utterance. Note that due to the increase in processing time when considering all actions up to the current time, we use the world state and its differences instead of all actions. ΔW_i is further divided into those representing interlocutor A's changes ΔW_i^A and interlocutor B's changes ΔW_i^B. The world state W and changes in the world state ΔW are represented in 10 × 10 × 4 voxels. W contains the block IDs at each coordinate, while ΔW stores the block IDs after the changes (if there is no change, it is 0).

For the actions taken in the most recent N turns, we include action types (UTT, BLOCK, SKIP, and FINISH) and, in the case of UTT or BLOCK, we also include the content of these actions. Each action type corresponds to a single token. In the case of UTT, we include the utterance text u_k. In the case of BLOCK, we include the change in the world state, denoted as Δw_k, occurring between the previous action a_{k-1} and the next action a_k. Δw is a compressed representation of block operations and is in the same format as ΔW.

The avatar's position (x_t, y_t, z_t) and orientation $(yaw_t, pitch_t)$ represent the coordinates and facial orientations in the environment.

4.3 Model Output

The model's output is action a_t at turn t. The model first outputs a token representing each action type, followed by, for UTT, the utterance text u_t, and for BLOCK, the set of block operations b_t. In this paper, to reduce complexity, the block operations are handled by dividing them into groups with up to L block operations, and we set L as the average size of b in the corpus, which is four. Therefore, the maximum length of b_t is four. All processing is performed by the LM Head. To facilitate this, tokens related to action types and block operations are added to the tokenizer in advance.

5 Experiment

5.1 Settings

In this study, we investigated the performance of the proposed model (Action-Utterance Model; AU) and baselines for the Next Action-Utterance Generation Task using the Collaborative Garden Task Corpus (Ichikawa and Higashinaka, 2022). Out of the 1,092 dialogues in the corpus, we randomly split the data and used 980 dialogues for training, 56 for validation, and 56 for testing.

To examine whether the proposed model can determine the appropriate action type to take next in a given context and perform suitable actions and utterances, we prepared baselines for action type classification, action generation, and utterance generation.

For action type classification, we used the following two baselines:

Random Selects one of the four action types at random.

Majority Always predicts the action type that is most frequently observed in the training data.

For action generation, we used the following baseline:

Random On the basis of the current world state, one to four feasible block operations are randomly selected.

Additionally, we established a human upper bound. For this, one of the authors predicted the set of next block operations to be performed for 20 samples randomly extracted from the test data.

For utterance generation, we used the following baseline:

Utterance Generation Only (UG) Transformer decoder model trained only for the utterance generation task. The model predicts the next utterance on the basis of all preceding utterances.

We used OpenCALM-Large[2], a Japanese LLM that contains 830 million parameters, and conducted LoRA tuning using the PEFT library (Mangrulkar et al., 2022). We optimized the model using Maximum Likelihood Estimation (MLE). During the evaluation, we used a checkpoint with the smallest loss calculated using the validation data.

5.2 Evaluation

We prepared the following evaluation metrics for each task: action type classification, action generation, and utterance generation. All metrics were computed by comparing the ground truth data with the inference results and yield values between 0 and 1, with higher values indicating better performance.

Accuracy Accuracy based on the classification results for action types and ground truth data.

Macro-F1 Macro-average of F1 scores calculated from the classification results and ground truth data for each action type.

BLEU-1, BLEU-2 Average BLEU-1 and BLEU-2 (Papineni et al., 2002) scores calculated by using generated utterances and gold response. If the system fails to generate an utterance due to the system predicting a value other than UTT, the value will be 0.

Distinct-1 Distinct-1 (Li et al., 2016) calculated on the basis of uni-grams of words present in generated utterances.

Jaccard Jaccard index calculated for the generated set of block operations $\bar{b}$ and the ground truth data b using the following formula.

$$Jaccard = \frac{1}{N}\sum_{i=1}^{N}\frac{|\bar{b_i} \cap b_i|}{|\bar{b_i} \cup b_i|} \quad (2)$$

To allow for a more lenient evaluation, two other metric values were also computed by considering only the set of block operation types (**Jacc-type**) and only the set of block

[2] https://huggingface.co/cyberagent/open-calm-large

Model	Accuracy	Macro-F1
Random	0.24	0.19
Majority	0.61	0.19
AU (ours)	**0.81***	0.67

Table 1: Evaluation results for action type classification. Bold indicates best value. * indicates **Accuracy** was significantly better than Random and Majority at $p < 0.05$ in McNemar test with Bonferroni correction.

Model	Jaccard	Jacc-type	Jacc-loc
Random	0.00	0.04	0.00
AU (ours)	**0.17***	**0.38***	**0.27***
Human	0.30	0.55	0.32

Table 2: Evaluation results for action generation. Bold values indicate best value except for Human. * indicates metrics were significantly better than Random at $p < 0.05$ in Wilcoxon signed-rank test with Bonferroni correction.

positions (**Jacc-loc**). If the system fails to generate an action due to the system predicting a value other than BLOCK, the value will be 0.

5.3 Results

Table 1 shows the results for action type classification, Table 2 shows those for action generation, and Table 3 shows those for utterance generation. The proposed model significantly outperformed the baselines in action type classification and action generation. Furthermore, it achieved a higher score in utterance generation compared with the baseline, especially in terms of Distinct-1. These results show that the proposed model can effectively determine the appropriate next action types and generate better actions and utterances by handling both action and utterance generation in a unified model.

Figure 3 shows a sample from the test data and actions generated by the proposed model and the baseline. The proposed model, while selecting to utter, generated utterances relevant to the flow of the dialogue and the current world state.

6 Analysis

To understand the current challenges with the proposed model, we conducted a detailed analysis of the inference results. When categorizing action generation on the basis of the characteristics of ground truth block operations, we found that the Jaccard index was high at 0.20 when the same types of blocks as the previous actions were included, while it dropped significantly to 0.07 when they were not. Similarly, when adjacent blocks were included as previous actions, the Jaccard index was high at 0.21, but it was low at 0.10 when they were not. These results show that predicting cases unrelated to the last actions is a challenge. They also suggest that there is insufficient grounding between dialogue and the world state and a lack of understanding of symmetries and regularities that humans comprehend.

Model	BLEU-1	BLEU-2	Distinct-1
UG	0.148	0.096	0.136
AU (ours)	**0.153**	**0.098**	**0.155**

Table 3: Evaluation results for utterance generation. Bold indicates best value.

7 Conclusion

In this study, we proposed a novel model for simultaneously generating actions and utterances during collaborative work in Minecraft. The experimental results showed that the proposed model can autonomously determine whether to act or utter and generate better actions and utterances than the baselines. Furthermore, we analyzed the inference results and revealed the difficulty in generating actions unrelated to the last actions.

There are limitations in our study. We compared our proposed model to simple baselines for action type classification and action generation; we need to perform comparisons with models introduced in previous work such as (Mohanty et al., 2023) and (Mehta et al., 2023). In addition, we only conducted turn-level evaluations; we need to consider dialogue-level evaluations in order to more accurately measure the model's performance. While we utilized the Jaccard index as the evaluation metric for action generation in this paper, the similarity of block operations may not be sufficient; therefore, we would like to conduct human evaluations and explore more appropriate evaluation metrics.

Additionally, we will also work towards building systems capable of actual collaborative dialogue. Due to the high flexibility of the next action and collaborative work themselves, rather than optimizing by MLE, we will aim to acquire higher-performing dialogue agents by incorporating reinforcement learning.

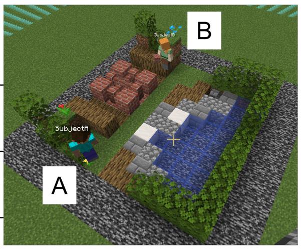
Last Actions	B: {(place, dandelion, -2, 4, 4)} A: 素晴らしい今までで一番出来栄えがいいです *(It's the best I've ever done, fantastic!)* B: そうですか *(Yes.)* A: SKIP B: よかった *(I'm glad to hear that!)* 
Worldstate	[12, 12, 7, 12, 14, 12, 7, 12, 12, 12, … , 1, 1, 1, 1, 1, 1, 1, 1, 1] (Corresponding to the state as in the right figure)
Gold Response	A: なんというか、別荘でのんびりしてる感じがします *(I feel like I'm relaxing at a vacation home.)*
UG	A: ありがとうございました *(Thank you.)*
AU (ours)	A: 木もいい味出してますね *(The wood also adds a nice touch, doesn't it?)*

Figure 3: Samples from test dataset and generation example for proposed model (translated to English by authors)

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19H05692.

References

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-Optimized Adapters for an End-to-End Task-Oriented Dialogue System. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7355–7369.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. 2021. Action-Based Conversations Dataset: A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3002–3017.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In *Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 18343–18362.

Jonathan Gray, Kavya Srinet, Yacine Jernite, Haonan Yu, Zhuoyuan Chen, Demi Guo, Siddharth Goyal, C. Lawrence Zitnick, and Arthur Szlam. 2019. Craftassist: A Framework for Dialogue-enabled Interactive Agents. *arXiv preprint arXiv:1907.08584*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

Takuma Ichikawa and Ryuichiro Higashinaka. 2022. Analysis of Dialogue in Human-Human Collaboration in Minecraft. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4051–4059.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. 2020. Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2589–2602.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. 2022. IGLU 2022: Interactive Grounded Language Understanding in a Collaborative Environment at NeurIPS 2022. *arXiv preprint arXiv:2205.13771*.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2013. A Data-driven Model for Timing Feedback in a Map Task Dialogue System. In *Proceedings of the 14th Annual Meeting of the Special*

Interest Group on Discourse and Dialogue-SIGdial, pages 375–383.

Nikhil Mehta, Milagro Teruel, Patricio Figueroa Sanz, Xin Deng, Ahmed Hassan Awadallah, and Julia Kiseleva. 2023. Improving Grounded Language Understanding in a Collaborative Environment by Interacting with Agents Through Help Feedback. *arXiv preprint arXiv:2304.10750*.

Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming Human-Centered AI Collaboration: Redefining Embodied Agents Capabilities through Interactive Grounded Language Instructions. *arXiv preprint arXiv:2305.10783*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative Dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.

Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. 2020. Gamification Platform for Collecting Task-oriented Dialogue Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7084–7093.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Scott E. Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. 2022. A Generalist Agent. *arXiv preprint arXiv:2205.06175*.

Zhengxiang Shi, Jerome Ramos, To Eun Kim, Xi Wang, Hossein A. Rahmani, and Aldo Lipani. 2023. When and What to Ask Through World States and Text Instructions: IGLU NLP Challenge Solution. *arXiv preprint arXiv:2305.05754*.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.