

Manipulating Multi-Agent Navigation Task Via Emergent Communications

Han Yu¹, Wenjie Shen¹, Ling Huang², Caixia Yuan^{1*}

¹Beijing University of Posts and Telecommunications, Beijing 100876, China;

²AHI Fintech Limited, Fangshan Dist., Beijing 102400, China

yuhan2021@bupt.edu.cn, shenwj@bupt.edu.cn, linghuang@fintec.ai, yuancx@bupt.edu.cn

Abstract: This paper presents a new approach for multi-agent communication via emergent language in a navigation task. The task involves a tourist and a guide who communicate through an emerged language grounded in the environment to help the tourist reach a target place. The paper proposes a collaborative multi-agent reinforcement learning framework that enables the agents to generate and understand emergent language, with the goal of solving the task. We evaluate the results on a simulated navigation game with 3,000 scenarios and multi-turn dialogues. Results show that the proposed approach achieves competitive performance in both language understanding and task success rate. The paper also provides explanations of the emerged language by comparing the language patterns with the environment structures.

Keywords: Multi agent; Emergent language; Vision-language navigation

1 Introduction

Effective communication is crucial for multiple agents to cooperate. Recently, many works have focused on interactions between artificial agents and humans, while some research is being conducted on communication between artificial agents. However, most of these works have focused on single-turn communication with a unidirectional message pass, emphasizing properties of the emergent language such as compositionality and interpretability. Multi-turn conversation is more analogous to human language, where language generation and understanding should be mutual rather than unidirectional. Therefore, we propose a framework for agents to generate multi-turn dialogues involving two agents.

To demonstrate the feasibility of our framework, we introduce a new task adapted from the vision-language navigation (VLN) task in the human-machine communication field. The task requires a guide (agent B) to communicate with a tourist (agent A) and help it find the target location. Unlike traditional VLN tasks, both the tourist and the guide are machines in our setting. Moreover, we assume that the guide does not know the initial position of the tourist, so it must not only provide guidance but also confirm the tourist's location.

Our contributions can be summarized as follows:

1) We have investigated emergent language in the context of multi-turn conversations, and have designed a navigation task adapted from the vision-language navigation (VLN) task to provide a suitable scenario for studying this type of language.

2) Compared to methods that involve agents speaking natural language, our approach is more cost-effective as it does not require expensive annotations. As a result, our method is a more practical solution for facilitating communication between agents.

3) To the best of our knowledge, our work represents the first attempt to propose a vision-language navigation (VLN)-like task in a two-agent cooperation scenario. In addition, we have developed a benchmark for evaluating performance on this task, which provides a potential solution to the challenges of multi-turn communication between agents. Our contributions provide a novel and valuable framework for studying the emergent language in multi-turn conversations between agents.

2 Related work

2.1 Vision language navigation (VLN)

Navigation tasks that involve both vision and natural language information have gained significant attention in recent years. Typically, these tasks include a human, an artificial agent, and an environment. The agent communicates with human in natural language, receiving requests or instructions and seeking further guidance from the human as needed. The agent must navigate and interact with the environment to gather more information to complete the task, with human providing requests and guidance to assist the agent in task completion. These types of tasks have become increasingly important as they bridge the gap between natural language understanding and embodied perception, potentially enabling more natural and intuitive human-machine interactions.

The primary objective of vision-language navigation (VLN) tasks is to guide the agent to a target location that is known only to the human side. Anderson et al. [1] introduced a VLN dataset named Room-to-Room (R2R), which is based on Matterport3D [2] data, and provided a benchmark for this task. Chen et al. [3] proposed the Touchdown dataset, which applies the VLN task to outdoor environments. Additionally, some scenarios have been proposed where agents must interact with objects in

the environment. For instance, Narayan-Chen et al. [4] integrated the VLN task into the Minecraft game and instructed the agent to complete building tasks. The emergence of VLN tasks bridges the gap between natural language understanding and concrete perception, providing possibilities for achieving more natural and intuitive human-machine interaction. From a language format perspective, these tasks can be divided into three categories [5]:

1) Initial instructions: The agent is provided with initial instructions for the entire navigation episode. Some studies have presented highly detailed instructions [1-2, 6-7], such as "Walk down one flight of stairs and stop on the landing." In contrast, other studies have provided more coarse-grained instructions [8]. However, tasks that fall into this category only provide instructions at the beginning of the episode, which limits the generation of multi-turn communication between agents.

2) Oracle guidance: During the navigation process, the agent may request additional natural language guidance to obtain more information [9]. However, in most cases, the oracle can only respond with simple phrases such as "turn left" and "turn right."

3) Dialogue: Supervised dialogue is commonly utilized as a signal [10-11]. However, fixed dialogue can constrain the flexibility of the policy.

Applying the VLN task to multi-agent settings can alleviate or solve the issues mentioned above. To our knowledge, Talk the Walk [10] is the only existing work that has applied the VLN task to multiple agents and is similar to our task settings. However, Talk the Walk only focuses on the localization aspect, where the guide locates the tourist, overlooking the guidance aspect, in which the guide creates a route plan and provides guidance to the tourist, while the tourist takes the corresponding actions. To model this aspect, Talk the Walk employs a random walk and only implements a minimal baseline, rendering the VLN task with multi-agent settings incomplete.

2.2 Emergent language

Emergent language refers to a language that arises naturally without explicit instruction or pre-existing grammar rules. It emerges from the interaction of individuals in a community or system, who communicate with each other to achieve a shared goal or complete a task. Emergent languages can develop in various settings, such as in the emergence of pidgin languages from contact between different language speakers, or in the development of communication systems among animals or artificial agents. With regard to agent relationships, most research has focused on fully cooperative tasks, such as referential games, which were adapted from the Lewis signaling game by Lazaridou et al. [12]. In the referential game, there are two roles: speaker and listener. The speaker has to give instructions to the listener so that it can select the target picture from the candidates

successfully. Kalinowska et al. [13] and Kajic et al. [14] have introduced navigation tasks, but these are constrained to specific environments. In addition to fully cooperative tasks, such as referential games, some semi-cooperative tasks have been proposed in the literature, such as the negotiation game introduced by Cao et al. [15]. In this game, two agents must negotiate to achieve higher scores for themselves. Emergent language research has also investigated fully competitive scenarios, like the circular-based sender-receiver game developed by Noukhovitch et al. [16]. In this game, the sender tries to sway the receiver's decision on a target position using their own target position as a reference. Most studies on emergent language focus on single message exchanges between agents, with one agent generating language and the other interpreting it. However, some studies have examined multi-turn conversations, such as those by Cao et al. [15] and Evtimova et al. [17]. The former focuses on a semi-cooperative case, while the latter employs a bag-of-words approach that generates either 0 or 1, indicating the presence or absence of a word in the generated sentence. This bag-of-words approach does not account for the dependence between words, resulting in less coherent sentences.

3 Navigation task

Our proposed task is inspired by a real-world scenario where two agents, agents A and B, cooperate to help agent A find a target location. As depicted in Figure 1, agent A is located in an environment and only able to see the surroundings around it, while agent B possesses the floor plan and knowledge about the target location, but does not know agent A's whereabouts. The objective of the game is for the agent B to tell its surroundings to A, and A to tell B how to move forward. The process is repeated until A reaches the destination, or the communication meets the predefined maximum rounds. If objective is not achieved, the task is considered a failure.

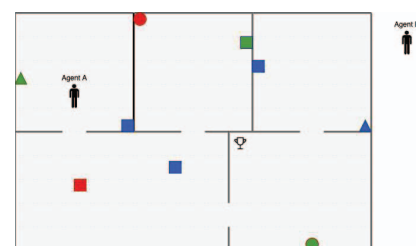


Figure 1 The simulated navigation task

The problem is divided into three steps and formalized as follows:

A set of K candidate houses is considered, each of which contains a collection of rooms $R = \{r_0, r_1, \dots, r_N\}$. At each iteration, one house is randomly selected. Subsequently, two rooms are sampled from the chosen house, with one designated as the target room r_o , and the other as the starting room r_s . Each room is equipped

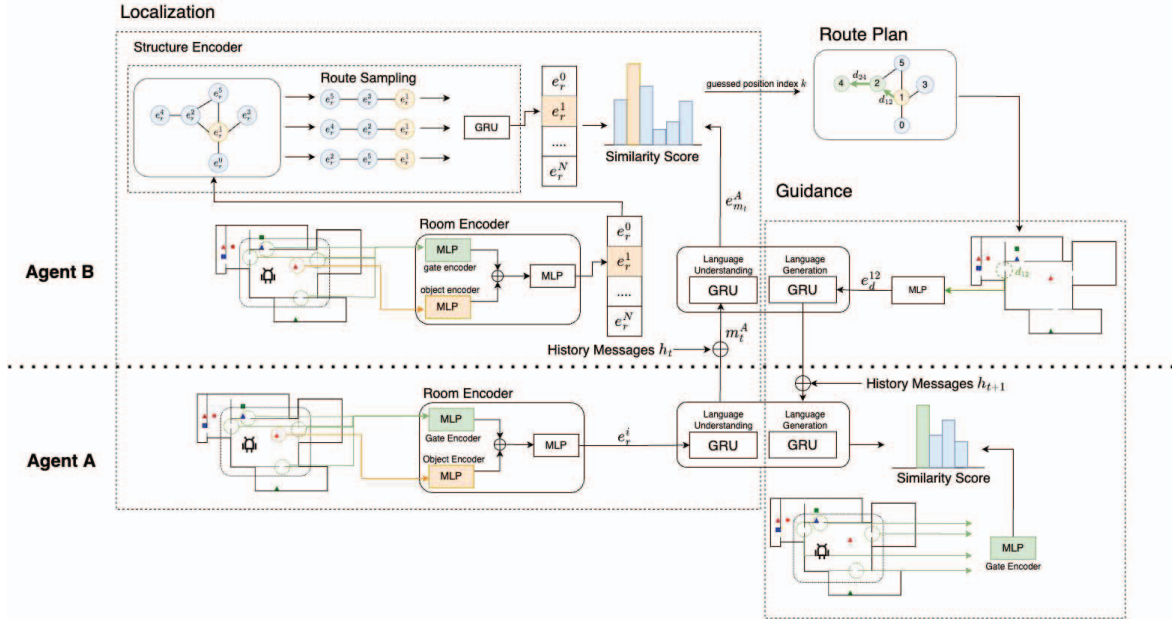


Figure 2 Model structure

with a set of gates $D_i = \{d_{ij}, j \in \text{Neighbor}(i)\}$, and a graph $G = (R, D)$ is constructed for each house to represent the spatial relationships between rooms. The task involves two agents, A and B, each with their own parameters θ_A, θ_B , and their input-output specifications depend on the specific task.

(1) In the localization task, agent A takes is the vector of the room and gives out the messages of step t consisting of M tokens $m_t = \{v_0, v_1, \dots, v_M\}, v_i \in V$. While agent B collects history dialogue $h_t = (m_0, m_1, \dots, m_t)$ and tries to guess where A is.

(2) In the route plan task, agent B uses a rule-based approach to plan the route. In the training process, agent B get the next step A should take according to the actual position r_s and the target position r_o . While during the testing process, agent B get the next step according the guessed position of A $\arg \max(\theta_B(h_t, G))$ instead of actual position.

(3) The only difference between two tasks is that in location task agent A is speaker and B is listener, while in guidance task agent A is listener and B is speaker.

4 Model

The model used in the project includes an encoder encoding the surroundings of A, a route encoder embedding the neighbor information by sampling routes ended with the target room, a language generation module in agent A implemented with GRU, and a language understanding module in agent B used for guessing the position of A.

Figure 2 shows the model structure for the localization task. Firstly, the information about the house containing features of a set of rooms $\{r_0, r_1, \dots, r_N\}$ is given. Agent A encodes the target room information r_o including objects and gates and then uses the encoded vector to

generate corresponding emergent language with its language generation module $e_r^o = GRU(MLP(r_o))$. With the message coming from agent A in this turn, and the history sentences from A, agent B makes its guess through the topological structure of rooms $\theta_B(h_t, G)$. In this topological graph, every room is represented with a node, and every door connecting neighbor rooms is represented with an edge between these rooms. For node k , we sample T routes ended with this node route j^k , which have the same length of history movement length t . Node k has representation $e_s^k = \frac{1}{T} \sum_i^T GRU(\text{route}_i^k)$ including topological structure information.

Meanwhile, B encodes the history sentences with another GRU and computes the similarity between this encoded history e_{m_t} with each node representation e_s^i calculated above to find the most possible position of agent A. Similarly, the model for the guidance task also includes an encoder, a language generation, and a language understanding module for giving instructions to agent A about its next step.

Considering the non-differentiability of the discrete messages passing between agents, those models are trained with REINFORCE algorithm. We use the Floyd algorithm for the route plan task to find the shortest paths. s represents the current state of the agent. v means the chosen vocabulary. r and d refer to the predicted room and door, respectively. Losses in localization and guidance tasks can be calculated with the following equations:

$$\begin{aligned} \nabla \theta_A J_{\text{localization}} &= E \left[\frac{1}{M} R(\cdot) \sum_i^M \nabla \theta_A \log \pi_A(v_i | s_A) \right] \\ \nabla \theta_B J_{\text{localization}} &= E [R(\cdot) \nabla \theta_B \log \pi_B(r | s_B)] \\ \nabla \theta_B J_{\text{guidance}} &= E \left[\frac{1}{M} R(\cdot) \sum_i^M \nabla \theta_B \log \pi_B(v_i | s_B) \right] \\ \nabla \theta_A J_{\text{guidance}} &= E [R(\cdot) \nabla \theta_A \log \pi_A(d | s_A)] \end{aligned} \quad (1)$$

Table I Comparison between discrete VAE and our model

Model	Task	Accuracy	ρ	Purity
Continuous	Localization	0.96	-	-
Continuous	Navigation	0.89	-	-
Discrete VAE	Localization	0.81	-0.0113	0.556/0.667/0.148
Ours	Localization	0.84	0.4775	0.407/0.457/ 0.235
Discrete VAE	Navigation	0.69	-	-
Ours	Navigation	0.86	-	-

Finally, we put those sub-tasks together to complete the navigation task. In the training process, the input of the localization part is the surroundings of agent A, and the output of it is the guessing position made by agent B. However, the input of the route plan part in the next timestep is the actual position of A, instead of the guessing one, preventing the error from localization influencing other modules. While in the testing process, the input of the route plan module comes from the last localization module.

5 Experiment

5.1 Dataset

We constructed the dataset with 6 rooms for each navigation task, with every object in the room having 3 optional colors and 3 optional shapes. In addition, some features not directly relevant to the task were included, such as the gate color, which we assumed to be the same across all rooms in the same house.

5.2 Comparison model

Rolfe [18] proposed a discrete VAE that addresses the challenge of backpropagation through discrete variables, and successfully constructs a semantic representation of images.

In this study, we utilize a discrete VAE as a comparison model by replacing the room encoder and gate encoder, along with the language generation component. First, we train the discrete VAE with the task reconstructing the rooms and gates. Then, we fix the discrete VAE as the room encoder and gate encoder and train the structure encoder and language modules of both agents including the language understanding and generation modules.

5.3 Results

Figure 3 and Figure 4 depict the changes in accuracy on the validation set and the rewards over the epochs respectively. In this experiment, the length of messages exchanged between agents A and B is set to 3, and there are 20 candidate tokens for the localization task and 40 candidate tokens for the navigation task. It is noteworthy that, since the model is trained from scratch, there is no prior meaning for the candidate tokens.

Table I outlines the results of the comparisons of discrete

VAE [18] and our model. It should be noted that the adapted discrete VAE already has the initial meaning of each token before training on the specific tasks. Considering that the presentations of room and routes are trained based on the reconstruction task in the discrete VAE model, making the discrete presentations contain more redundant information. With the limitations of the message length and vocabulary size in discrete communication scenarios, discrete VAE may communicate less useful information compared with our model. We also show the case when agents can pass continuous messages, giving us a ceiling performance when the communication channel is not restricted.

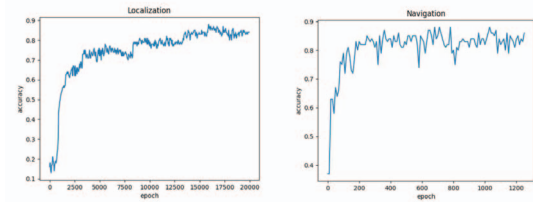


Figure 3 The change of accuracy over epoch in different tasks
(a) localization task (b) navigation task



Figure 4 The reward changes of the training set and the validation set

To evaluate the generated language, we use topographic similarity and purity. It is supposed that a good language tends to give similar expressions describing a similar object. Therefore, we use topographic similarity ρ as the metric, which is the Spearman correlation between the message similarity sequence and the room vector similarity sequence. Table I illustrates the comparison between discrete VAE and our model. From Table I, we observe that our method outperforms the discrete VAE largely. To make the emergent language more interpretable and give more hints about the information it

encodes, we calculate the purity score with the following procedure: Firstly, we only change the color, shape, and direction attributes of the object and fix other features. Then, we cluster those rooms with the generated message. Labeled these rooms with their object color, shape and direction separately, we get the corresponding purity. And the purity score can tell us the information that the emergent language encoded. From the Table I, we can tell that compared with discrete VAE, our model focuses more on the direction and less on color or shape. It is reasonable because there are only 3 choices for color and shape which makes it less useful for disguising the target room from others, comparing with the direction with 8 possible values.

In addition, we investigate factors that may impact task performance in Figure 5. Contrary to our initial intuition that longer sentences would lead to greater performance due to increased expressive power, we observe that localization accuracy initially increases but then drops as the sentence length increases. This is likely because when the sentence length is sufficient for effective communication, an excessive number of candidate tokens may actually increase the action space and makes the training process more difficult. We also find that the shortest route length between start and target positions can impact task performance, although the performance drop is not exponential as the shortest route length increases. This result shows that the task performance is in downward trend in total as the task difficulty increase.

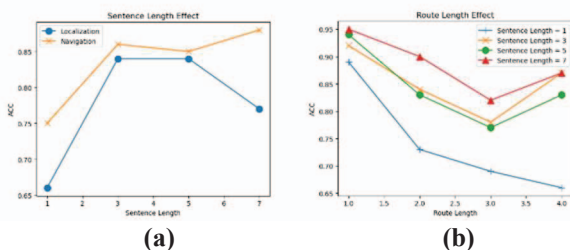


Figure 5 The effect of (a) sentence length and (b) the shortest route length between the start and target on the task performance

6 Conclusions

In this paper, we present a novel approach to the vision-language navigation (VLN) task by introducing multi-agent communication. The proposed navigation task is adapted from the human-machine communication scenario and is designed to enhance the flexibility of the communication policy and generate more complex dialogue formats compared to traditional VLN tasks. The experimental results demonstrate the feasibility of applying multi-agent communication to the VLN task and provide a comprehensive benchmark for evaluating this approach. Overall, this study contributes to promote the research of VLN by introducing a new perspective on communication.

References

[1] Anderson P, Wu Q, Teney D, et al. Vision-and-language

navigation: Interpreting visually-grounded navigation instructions in real environments. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018: 3674-3683.

[2] Chang A, Dai A, Funkhouser T, et al. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

[3] Chen H, Suhr A, Misra D, et al. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019: 12538-12547.

[4] Narayan-Chen A, Jayannavar P, Hockenmaier J. Collaborative dialogue in Minecraft. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019: 5405-5415.

[5] Gu J, Stefani E, Wu Q, et al. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

[6] Jain V, Magalhaes G, Ku A, et al. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019.

[7] Krantz J, Wijmans E, Majumdar A, et al. Beyond the nav-graph: Vision-and-language navigation in continuous environments. *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII* 16. Springer International Publishing, 2020: 104-120.

[8] Zhu F, Liang X, Zhu Y, et al. Soon: Scenario oriented object navigation with graph-based exploration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021: 12689-12699.

[9] Chi T C, Shen M, Eric M, et al. Just ask: An interactive learning framework for vision and language navigation. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, 34(03): 2459-2466.

[10] De Vries H, Shuster K, Batra D, et al. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*, 2018.

[11] Thomason J, Murray M, Cakmak M, et al. Vision-and-dialog navigation. *Conference on Robot Learning*. PMLR, 2020: 394-406.

[12] Lazaridou A, Hermann K M, Tuyls K, et al. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.

[13] Kalinowska A, Davoodi E, Mathewson K W, et al. Towards Situated Communication in Multi-Step Interactions: Time is a Key Pressure in Communication Emergence. *Proceedings of the Annual Meeting of the Cognitive Science Society*. 2022, 44(44).

[14] Kajić I, Aygün E, Precup D. Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint arXiv:2004.01097*, 2020.

[15] Cao K, Lazaridou A, Lanctot M, et al. Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980*, 2018.

[16] Noukhovitch M, LaCroix T, Lazaridou A, et al. Emergent communication under competition. *arXiv preprint arXiv:2101.10276*, 2021.

[17] Evtimova K, Drozdov A, Kiela D, et al. Emergent communication in a multi-modal, multi-step referential game. *arXiv preprint arXiv:1705.10369*, 2017.

[18] Rolfe J T. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.