An Experimental Study of Emergence of Communication of Reinforcement Learning Agents

Qiong Huang(✉) and Doya Kenji(✉)

Okinawa Institute of Science and Technology Graduate University,
Onna, Okinawa 904-0495, Japan
{qiong.huang,doya}@oist.jp

Abstract. Ability to use language is an essential requirement for human-level intelligence. For artificial general intelligence, the ability to learn and to create language is even more important [1]. Most previous models of learning and emergence of language took successful communication itself as the task target. However, language, or communication in general, should have evolved to improve certain fitness of the population of agents. Here we consider whether and how a population of reinforcement learning agents can learn to send signals and to respond to signals for the sake of maximizing their own rewards. We take a communication game tested in human subjects [2,3,6], in which the aim of the game is for two players to meet together without knowing exact location of the other. In our decentralized reinforcement learning framework with communicative and physical actions [4], we tested how the number N of usable symbols affects whether the meeting task is successfully achieved and what kind of signaling and responding are learned. Even though $N = 2$ symbols are theoretically sufficient, the success rate was only 1 to 2%. With $N = 3$ symbols, success rate was more than 60% and three different signaling strategies were observed. The results indicate the importance of redundancy in signaling degrees of freedom and that a variety of signaling conventions can emerge in populations of simple independent reinforcement learning agents.

Keywords: Multi-agent system · Reinforcement learning · Communication · Meeting task

1 Introduction

While communication is ubiquitous among animals and plants, unique features of human language are that the mapping between the signals and meanings, or appropriate responses, is not genetically fixed but learned by each individual and that a variety of vocabularies and syntactic conventions emerge through cultural evolution in different populations. How such capability is realized by evolution and learning [1] is a major question in artificial general intelligence.

P. Hammer et al. (Eds.): AGI 2019, LNAI 11654, pp. 91–100, 2019.
https://doi.org/10.1007/978-3-030-27005-6_9

Many models of learning and emergence of language have been proposed, but most of them took successful communication itself as the task target. From evolutionary viewpoint, however, language, or communication in general, should have emerged for the sake (or outcome) of improving certain fitness of the population of agents. Here we consider whether and how a population of reinforcement learning (RL) [9] agents can learn to send signals and to respond to signals for the sake of maximizing their own rewards.

Previous research [8] examined whether and how a simple form of communication emerges between reinforcement learning agents in an intrusion game, where positive reward is acquired by stepping into the other's territory while negative reward is incurred by collision of the two. Agents with light signaling capability learned a variety of policies of signaling and responding to signals to realize coordination, dominance, and complex behaviors. Another study in [5] introduced goal-directed utterance selection through learned internal models of how others respond to utterance. Agents were able to decide when to use language or not and select the appropriate utterance to achieve their specific goals of obtaining a certain type of food. More recently, Mordatch and Abbeel [7] demonstrated grounded compositional languages can emerge among deep reinforcement learning agents. In this work, however, all agents shared the same policy, which is closer to genetically shared communication scheme like in bees, rather than to human language learned independently by each individual.

Galantucci [2] developed a "meeting" task in which a pair of human participants learn to exchange graphic symbols to meet in the same room while not explicitly knowing the other's location. Different pairs converged to different conventions of what each symbol means and how to respond to them. Konno and colleagues [6] reproduced the experiment [2] in and further proposed a learning model based on the Adaptive Control of Thought-Rational (ACT-R) architecture.

In this paper, we propose a decentralized multi-agent RL framework for signaling and physical actions and test its performance with the meeting task [2,6]. Our previous work [4] tested the framework in a simpler task where agents get a reward by simultaneously arriving a fixed position in a grid-world. We test how the number N of usable symbols affects whether the meeting task is successfully achieved and what kind of signaling and responding are learned. We show that simple RL agents are able to create different signaling patterns for communication to guide the actions of each other. Analysis of the learned policies of signaling and corresponding movements shows that a variety of "meanings" can emerge through interactions of RL agents.

2 Methods

2.1 Learning Framework

We proposed a split-Q state-action function framework [4] which separates the physical moving actions and communication signal actions. The learning framework could be explained with Fig. 1. In this framework, each agent possesses two state-action function pairs which are for state-moving action and state-signaling (Q_i^p and Q_i^c, where $i = 1, 2$), respectively. In each episode, the agent will select a signaling action (i.e. a message that will be shared with the other agent), send to the other agent and take a physical movement afterwards. The environment would therefore generate reward and the next states which would feedback to both agents.

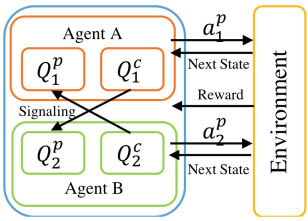

Fig. 1. Learning framework.

2.2 Problem Formulation

The setting of the experiment is a "meeting" task based on [2,6], which could be viewed as a Partially Observable Markov game. In the present study, we utilized two reinforcement learning agents to see whether they could learn the meaning of the signaling from each other to enter the same slot within only one step move. The game field is a 4-rooms environment (Fig. 2). Agents are initialized in different rooms and have no prior knowledge of where the other agent is. Each agent possesses 5 possible movements (up, down, left, right, and stand), and several numbers of signals that they could utilize to send to the other agent. Agents can choose which signal they would like to use to send to the other agent before the move and will receive a score after each movement. Moreover, no predefined meanings are postulated for the signals. If they met in the same room after one movement, both agents receive a reward = 2; otherwise, they will both receive a reward = −1. Once they met in the same room or reached maximum steps in per round, they will be reset with new random starting positions, i.e., different separate rooms. To simplify the symbols, here we used numeric numbers to represent the signals. The goals are to explore whether agents can learn to move to meet in the same room within one move after learning, how agents utilize the signals, and what kind of factors are influencing this process.

In this environment, we assumed that the two learning agents ($i = 1, 2$) each has its own physical state set S_i^p, a communication state set S_i^c, a moving action set A_i^p, and a communication action set A_i^c. Each agent i has two state-action pair value functions which stand for the moving and communication respectively. One state-action pair function is $Q_i^p(s_i^p, s_i^c, a_i^p)$ which evaluates a moving action $a_i^p \in A_i^p$ at a physical state $s_i^p \in S_i^p$ and a communication state $s_i^c \in S_i^c$. The other

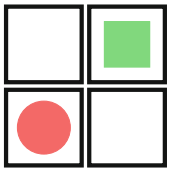

Fig. 2. Game field. Circle and rectangle denotes different agents.

state-action pair function is $Q_i^c(s_i^p, a_i^c)$ which evaluates a communication signal $a_i^c \in A_i^c$ at a physical state s_i^p. The communication action of the agent therefore changes the communication state of the other agent. The communication state satisfies the following formula $S_i^c = f(a_j^c) = a_j^c$ where i and j denotes different agent.

Actions are selected from the following conditional probability functions: $\pi_i^p(a_i^p \mid s_i^p, s_i^c) \propto \frac{e^{Q_i^p(s_i^p, s_i^c, a_i^p)/\tau_i^p}}{\sum_b e^{Q(s_i^p, s_i^c, b)/\tau_i^p}}$, and $\pi_i^c(a_i^c \mid s_i^p) \propto \frac{e^{(Q_i^c(s_i^p, a_i^c)/\tau_i^c}}{\sum_b e^{Q(s_i^p, b)\tau_i^c}}$, where τ_i^p and τ_i^c are temperatures which control randomness. Algorithm 1 describes how the learning rules are updated.

Algorithm 1. Updating rule with split Q-learning framework.

Initialize $Q_i^p(s_i^p, s_i^c, a_i^p)$ and $Q_i^c(s_i^p, a_i^c)$ arbitrarily
repeat
  for all agents i **do**
  Initialize s_i^p
  take a_i^c **update** $s_i^c = a_\ell^c$ $(\ell \neq i)$
  repeat
    choose a_i^p, observe new states $s_i'^p$,and r_i
  for all agents i **update**
    $Q_i^p(s_i^p, s_i^c, a_i^p) \leftarrow (1-\alpha)Q_i^p(s_i^p, s_i^c, a_i^p) + \alpha(r_i + \gamma max_b Q_i^p(s_i'^p, s_i'^c, b))$,
    $Q_i^c(s_i^p, a_i^c) \leftarrow (1-\alpha)Q_i^c(s_i^p, a_i^p) + \alpha(r_i + \gamma max_b Q_i^c(s_i'^p, b))$,
    $s_i^p \leftarrow s_i'^p$, $s_i^c \leftarrow s_i'^c$
until Termination Conditions
where α is the learning rate and γ is the discount factor

3 Result and Discussion

3.1 Cases with Different Number of Signals

In our experiments, number of symbols N varies from 2 to 5. And we performed 100 groups of agents with 10,000 runs for each group and repeat for 5 times. Each agent also has separated learning rate α_i^p, α_i^c and temperature τ_i^p, τ_i^c. Here the incremental in the inverse temperature with an annealing equation follows $\tau_i^{p,c} = 1/(1 + \tau_i^{p,c} \times epi)$, where epi is the number of learning episode.

(i) $N = 2$

When agents have 2 symbols to choose to forward to the other agent, there are very rare cases that they can learn to meet in the same room after communication. Among all 100 groups of agents, there were only 1 or 2 groups which succeed in meeting each other within one step move. The signaling that agents learned to use (while there are only 2 possible signals, we used green and red color to represent the utilization of the signals) and signaling pattern paired with the successful case (the optimal solution) are listed in Fig. 3. The motion policies are colored in the same color of signals received from the other agent. For agent A, the signal it received in one room is as the pattern on the right side of Fig. 3 (i.e., agent B's signaling) and its neighborhood signal is the same whichever room it starts. For agent A, when agent A received the same signal from its neighborhood rooms, it chose actions which allow it to stay in its current position; when agent A received the other signal, it will take other actions to move to an adjacent room. On the other hand, for agent B, the signal it received is as the pattern on the left side of Fig. 3 (i.e., agent A's signaling) and its neighborhood signal is always different. As we can see from Fig. 3, when agent B receives the same signal in one position, it chose actions to move towards the same side of the signals. In this case, agent A's signaling guides the action of agent B and tells it the room it shall goes to. Noticing that agent has three possible next state based on its current position, when it received the same signal based on its current position, it would take the same action.

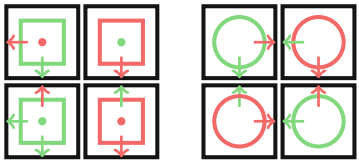

Fig. 3. Signaling pattern and motion policy pairs agents learned in successful groups when $N = 2$. Agent A and agent B is represented in box and circle, respectively. The signaling policies are showed by the edge color of the agent, and the motion policies are showed by arrow/dot for the received signal.

(ii) $N = 3$

When agents have 3 optional signals they could use for communication, compared with case (i), they have much higher probabilities to learn to enter the same room within one step move, and their signaling patterns also possess more diversity. With meta-parameter tuning, we found an optimal group of parameters $\alpha^p = 0.28$, $\alpha^c = 0.24$, $\tau^p = 0.0012$, and $\tau^c = 0.0021$. The probability of successfully meet in the same room is $63.4 \pm 2.7\%$. There are mainly 3 groups of combined signaling patterns pairs from the overall successful groups, and a visualized signaling pattern is represented in Fig. 4 as follows:

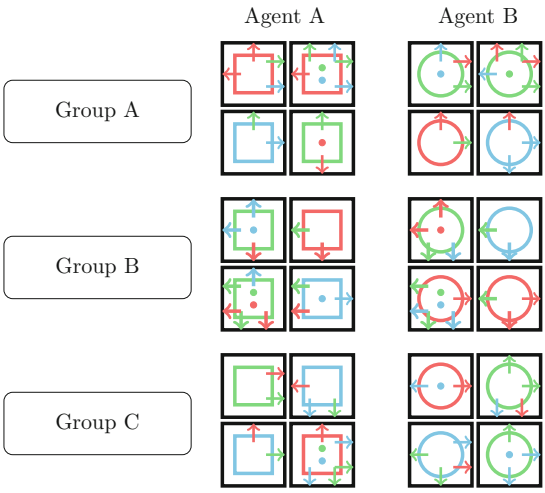

Fig. 4. Signaling patterns of agent A and agent B when $N = 3$, where different colors (green, cyan and red) represent for different signals agent learned after experimentation. Examples of the motion policies are showed by arrow/dot for the received signal. (Color figure online)

- Group A: agents show the same signal on the same side of the game field.
- Group B: agents show the same signal on one side of the game field and have one grid overlap of the same signal.
- Group C: one agents show the same signal in diagonal rooms of the game field.

Table 1 shows the number of each category (group). Group A and B are two most common signaling patterns in this case.

The moving policies that agents learned at the end in the well-learned groups also comply with certain rules. If one agent receives one specific signal, it will move accordingly to this signal, and vice versa for the other agent. In this situation, the signal received from the other agent guides the action of the present agent. As the agents are not in the same grid at the beginning, there exists an optimal solution in this environment when the number of signals is 3. Figure 5 is an example of successful communication policy learned by agent A and agent B in this premise, in which signaling patterns of Group B pairs are listed. As shown in Fig. 5, all 12 possible starting positions conditions are listed with their learned moving policies. For example, column 5–6 and line 2–3 shows the case that agent A sent the same message (colored in green) in the upper and bottom left rooms, and the message sending to agent B directs the direction in which it shall move (moving to the left). While agent A's moving action alters according

Table 1. Numbers of different groups when $N = 3$.

	Groups		
	Group A	Group B	Group C
Numbers	26	34	5

to the message it apprehends from the other agent in this 2 rooms, eventually they could meet in the same room under this signaling pattern within one-step move.

(iii) $N = 4$

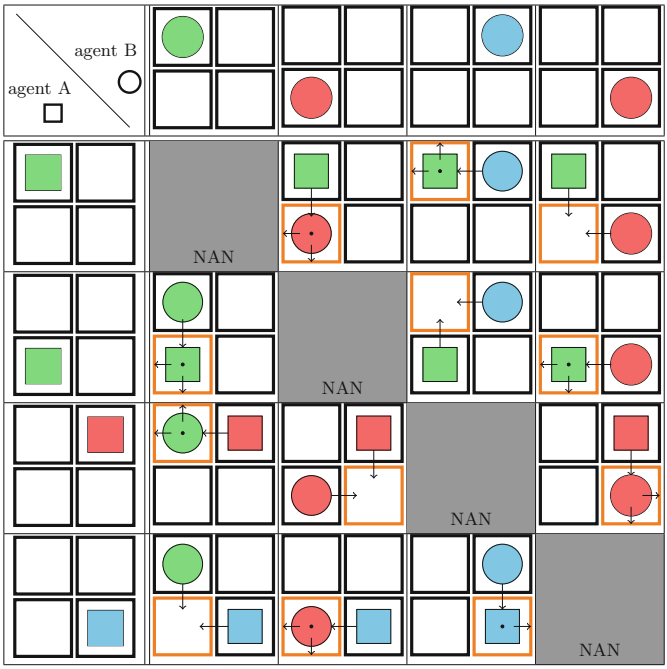

Fig. 5. One example of moving policy between agents when $N = 3$ (Group B). Three signals are colored in green, red and cyan, respectively. Orange box is the room agents meet after one step move. Arrows represent the directions agents move to, and dot is stand action. (Color figure online)

In this case, agents have 4 potential signals they could choose for communication, and the probability for agents to enter the same room within one step move keeps increasing compared with case (ii). Likewise, there are more patterns of signaling. With meta-parameter tuning, we found a group of optimal parameters where $\alpha^p = 0.3$, $\alpha^c = 0.25$, $\tau^p = 0.001$, and $\tau^c = 0.002$. The probability

of successfully meet in the same room is $86 \pm 2.2\%$. Here we list one experimental result where the number of successful groups is 87 (out of 100 groups), and patterns of the signaling are in 6 categories as follows:

- A1: Using pure 4 signals (e.g., 2, 3, 1, 4 in four rooms).
- B1: Using all 4 signals but have ambiguity in one position (e.g., 3, [1,4], 2, 1 in four rooms).
- B2: Using all 4 signals, have ambiguity, and 2 signals from the rest 3 positions (e.g., 4, [1,2], 4, 3 in four rooms).
- C1: Using all 4 signals and have ambiguity in two positions (e.g., [1,3], [1,3], 2, 4 in four rooms).
- D1: Using pure 3 signals (e.g., 2, 1, 3, 3 in four rooms).
- E1: Using 3 signals and have ambiguity in one positions (e.g., [1,4], 3, 3, 4 in four rooms).

The numbers of different groups are shown in Fig. 6. As agents have more choices of signaling ($N = 4$), agents not only shows the similar signaling pattern as they emerged in case (iii), but also have more variety of choices of signaling they could use. For example, in pattern A1, agents use 4 signals to represents different rooms, which is similar to a pattern found in [2] as human pairs.

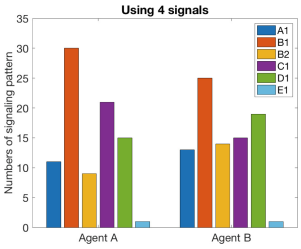

Fig. 6. The signaling pattern overall 87 successful learned groups.

(iv) $N = 5$

Similar results occurred when numbers of signaling is 5 compared with 4 signals. With meta-parameter tuning, we further found an optimal group of parameters where $\alpha^p = 0.25$, $\alpha^c = 0.28$, $\tau^p = 0.0011$, and $\tau^c = 0.0023$.

The probability of successfully meeting in the same room is $86.2 \pm 2.3\%$. In addition, besides the same signaling pattern as listed above in case (iii), there is another category F1 where agents uses 5 signals and have ambiguity in at least one room. Figure 7 shows the number of different patterns.

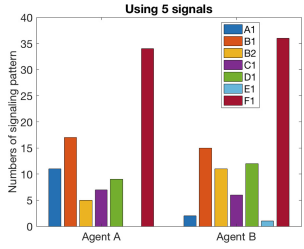

Fig. 7. The signaling pattern overall 83 successful learned groups.

3.2 Meta-Parameter Tuning

During implementation, we further noticed that the learning rate α and temperature τ are the two most contributing factors among other parameters which significantly influence the performance of the adaptive learners. We carried out grid

search with learning rate (from 0.10 to 0.30 with step length of 0.01) and temperature (from 0.0010 to 0.0030 with step length of 0.0001) for optimal parameters in this experiment, and we further noticed that both temperatures for moving and communication are not supposed annealing to 0 at the end of each runs, otherwise the success rate will decline significantly.

4 Conclusion

This paper presented an experimental study of the emergence of communication in decentralized reinforcement learning agents with a split-Q learning framework. By the learning framework, RL agents could learn to use signals to enter the same spot in one-step move under message exchange without knowing the position of the other agent, as human participants did in [2,6]. We found that meta-parameter tuning was crucial for descrying the optimal parameters for learning, in which learning rate α and temperature τ are the two contributing factors in determining agents' optimal behavior.

The minimal solution with $N = 2$ was rare found, possibly because the actions are highly depended on the positions of the agent. In [2], a common pattern for signaling found with $N = 4$ in experimental subjects were to use different symbols to represent different positions. In our experiment with $N \geq 4$, besides the clean coding A1, other patterns also emerged for agents to achieve an optimal behavior.

Future work will extend the framework to scenarios where emergence of compositional coding of multiple states, goals, actions, and objects can be investigated, as in [7]. Exploration of the roles of internal models [5] and the theory of mind is also an important future direction.

Acknowledgements. This work was supported by Ministry of Education, Culture, Sports, Science, and Technology KAKENHI Grants 23120007 and 16H06563, and research support of Okinawa Institute of Science and Technology Graduate University to KD.

References

1. Doya, K., Taniguchi, T.: Toward evolutionary and developmental intelligence. Curr. Opin. Behav. Sci. **29**, 91–96 (2019)
2. Galantucci, B.: An experimental study of the emergence of human communication systems. Cogn. Sci. **29**(5), 737–67 (2005)
3. Galantucci, B., Steels, L.: The emergence of embodied communication in artificial agents and humans. In: Wachsmuth, I., Lenzen, M., Knoblich, G. (eds.) Embodied Communication in Humans and Machines, chap. 11, pp. 229–256. Oxford University Press, Oxford (2008)
4. Huang, Q., Uchibe, E., Doya, K.: Emergence of communication among reinforcement learning agents under coordination environment. In: The Sixth Joint IEEE International Conference Developmental Learning and Epigenetic Robotics, pp. 57–58 (2016)

5. Klein, M., Kamp, H., Palm, G., Doya, K.: A computational neural model of goal-directed utterance selection. Neural Netw. **23**(5), 592–606 (2010)
6. Konno, T., Morita, J., Hashimoto, T.: Symbol communication systems integrate implicit information in coordination tasks. In: Dietterich, T., Becker, S., Ghahramani, Z. (eds.) Advances in Cognitive Neurodynamics (III), pp. 453–459. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-4792-0_61
7. Mordatch, I., Abbeel, P.: Emergence of grounded compositional language in multi-agent populations. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
8. Sato, T., Uchibe, E., Doya, K.: Learning how, what, and whether to communicate: emergence of protocommunication in reinforcement learning agents. Artif. Life Robot. **12**, 70–74 (2008)
9. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)