Gradual emergence of communication in a multi-agent environment

Shinjiro Tensho*, Satoshi Maekawa†, Junichiro Yoshimoto*,‡, Tomohiro Shibata* and Shin Ishii*
*Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0192 JAPAN
†National Institute of Information and Communications Technology, 3-5 Hikaridai Seika, Soraku, Kyoto, 619-0289 JAPAN
‡Initial Research Project, Okinawa Institute of Science and Technology, 12-22 Suzaki, Gushikawa, Okinawa, 904-2234 JAPAN
E-mail: shinji-t@is.naist.jp

Abstract

Communication is one of the most important keys for agents in a multi-agent environment to find global or semi-optimum policies. It has been attracting not only biologists but also engineers to make communication emerged in compliance with a given problem. Such emergence, however, is very difficult to realize from the computational viewpoint because it requires both sender and receiver agents to acquire the corresponding functions for communication individually but concurrently; such an emergent system tends to be evolutionally unstable. This article presents a new mechanism for emergent communication in a competitive multi-agent system. The key point of our mechanism is to introduce an environmental event (or factor) that agents have to learn to cope with. The acquired function for the environmental factor, then, drives the emergence of communication. Simulation results show that our approach allows gradual emergence of communication, which makes the agents to acquire higher fitness as compared with the case without communication.

keywords: communication, competitive environment, gradual emergence, reinforcement learning, evolution

1 Introduction

In recent years, studies of communication in multi-agent systems have been popular because they are important especially when dealing with large-scale problems that are intractable by single-agent approaches; potential targets are rescue robot systems and Robocup soccer systems for example.

In the engineering field, primitive functions of communication, such as protocol and encoding/decoding schemes, are often manually designed and built into agents. This approach may be effective in a completely specified and stationary domain, but would not be able to cope with partially known and/or non-stationary problems. An adaptive mechanism to emerge communication has possibility to avoid this defect.

In this article, we define that "communication" between agent A and agent B is to maximize acquired reward or to minimize suffered risk when agent A takes an action a_a then agent B takes an action a_b in response to the agent A's action. We call action a_a and a_b "communicative action" for agent A and agent B, respectively.

It is generally difficult to emerge communication concurrently in a multi-agent environment because a communicative action often costs. In addition, it is not guaranteed that the optimality of communication between two agents is generalized in a multi-agent environment, in which various interactions among agents affect each other and make the problem ill-conditioned [1].

This article presents a new mechanism to acquire communication emergently in a multi-agent competitive environment. The key point is to introduce time-invariant danger as an environmental event (or factor) that agents have to learn to cope with. The acquired function for the environmental factor, then, drives the emergence of communication making agents to acquire higher fitness than in the case without communication.

2 Setup

2.1 Agent's model

Figure 1 shows the agent's model used in this article. Each agent can observe four types of information at an arbitrary time step: 1) whether a big sound exists in the environment, 2) whether the agent faces another agent, 3) whether the agent is faced with resources, and 4) whether the agent occupies any resources in his hand. These sensory inputs are denoted by binary variables $s_i \in \{0, 1\}$, where $s_i = 0$ ($s_i = 1$) indicates that

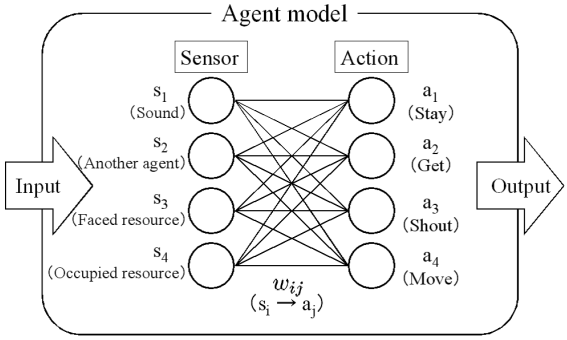

Figure 1: Agent model

the i-th information is true (false). For example, $s_1 = 1$ if the agent hears a big sound, and $s_4 = 0$ if the agent does not occupy any resource. After obtaining sensory inputs $S \equiv (s_1, \ldots, s_4)$, the agent takes one of four actions: 1) "*Stay*", 2) "*Get*", 3) "*Shout*" and 4) "*Move*". Here, "*Stay*" and "*Get*" are to keep the current situation and to pick up resources in front of it, respectively. "*Shout*" and "*Move*" are to raise a loud cry and to run away from the situation, respectively. Let a_i be the i-th action, e.g., a_1 indicates "*Stay*". Which action the agent takes for given sensory inputs is stochastic and the probability of a_i, π_j, is given by

$$\pi_j = \frac{e^{u_j}}{\sum_k e^{u_k}} \quad (1)$$

$$u_j = \sum_{i=1}^{4} w_{ij} s_i, \quad (2)$$

where $w_{ij}(i, j = 1, \ldots, 4)$ are connection weights that specify agent's behaviors.

Getting a resource increases agent's fitness to the environment. If an agent successfully obtains resources whose amount is denoted by r_t at time step t, the fitness f is increased by

$$f := f + r_t. \quad (3)$$

The aim of each agent is to maximize the fitness in its lifetime.

The colony of agents has two types of adaptation mechanism with different time scale [2]. One is the individual learning of connection weights based on the actor-critic method [3], and another is evolution. The learning process at time step t is as follows:

1. **Calculating TD-error** δ

$$\delta := r_t + \gamma V(S_t) - V(S_{t-1}) \quad (4)$$

where S_t denotes a sensory input at time step t and $V(S)$ is a value function for state S. $\gamma \in (0, 1)$ is a discount rate.

2. **Updating the value function** $V(S)$

$$V(S_{t-1}) := V(S_{t-1}) + \alpha\delta \quad (5)$$

where $\alpha \in (0, 1)$ is the learning rate.

3. **Updating connection weight** w_{ij}

$$w_{ij} := w_{ij} + \beta\delta s_i \quad (6)$$

where $\beta \in (0, 1)$ is a step size parameter.

After agents live fixed time steps, they move to the next generation. At the beginning of the new generation, N agents with the highest fitness at the end of the previous generation are duplicated whereas N agents with the lowest fitness are deleted. Accordingly, the population size in generations are fixed. To keep the diversity in agents, the non-deleted agents at the T-th generation are initialized as follows:

1. **Initializing fitness** f

$$f := 0 \quad (7)$$

2. **Initializing value function** $V(S)$

$$V(S) := 0, \quad \text{for all } S. \quad (8)$$

3. **Inheritance of the initial connection weight** w_{ij}

$$w^0_{ij}(T) = w^0_{ij}(T-1) + \epsilon, \quad (9)$$

where $w^0_{ij}(T)$ denotes the initial connection weight at the T-th generation. ϵ is a Gaussian noise.

4. **Initializing the initial connection**

$$w_{ij} = w^0_{ij}(T) \quad (10)$$

This genetic procedure called "Darwinism" [4] is the other type of adaptation, for the agents' colony, in our model.

2.2 Environment

We consider a multi-agent environment where all agents have an identical model explained in the previous subsection. Resources are distributed in the environment. If an agent takes a "*Stay*" action, nothing occurs and the fitness does not change. If an agent takes a "*Move*" action, the agent discards R_i resources in his hand and runs away to another situation. This results in decreasing the fitness by R_i. If an agent takes a "*Shout*" action, the fitness decreases by c_v because the action is assumed to require resources of c_v. If an agent takes "*Get*" when facing free resources, the agent gets them and the fitness increases by R. If an agent

takes "*Get*" when facing another agent that holds resources, a fight between these agents occurs and the former agent steals R resources from the latter agent. Since the fight costs d resources, the fitness of the former agent (the latter agent) increases by $R-d$ ($-R-d$). The immediate reward r_t at time step t corresponds to the increase or decrease in the fitness. If an agent takes "*Get*" when facing another agent that does not hold resources, a fight between these agents occurs but nobody wins. Since the fight costs d resources, the fitness of each agent just decreases by $-d$.

The objective of each agent is to maximize the fitness in its lifetime. If we assume that agents are able to know every information of the environment, it is easy to achieve their objective. Since our agent's sensors are not able to distinguish whether resources are free or occupied by another agent, however, an agent without any additional information may suffer from getting into fights by trying to acquire resources in front of it. Communication is one possible way to produce such additional information and hence to increase fitness by avoiding fights.

2.3 Mechanism

In our mechanism, the existence of dangers is the key to emerge communication, although the avoidance of the dangers are not directly related to communication. The dangers are distributed in the environment, and accompanied by a loud sound. Each agent probabilistically encounters a danger.

If an agent encounters a danger and does not avoid it, the agent suffers from big damage more than the reward through a fight between agents. Unless an agent acts "*Move*" when being faced with a danger, it receives a negative reward of $-d_d$ and the fitness decreases by $-d_d$. In this way, the sense of a danger naturally drives agents to associate it with running away as a policy. Note that an agent cannot recognize a danger directly, but a loud sound is an indirect signal associated with a danger. Thus, it is expected that all agents acquire the same policy for the case when they meet a danger.

Based on the above mechanism, in a competitive environment, if an agent is able to express the sense of a danger against other agents, it can avoid fighting and keep the current resources without any risk. Therefore, an action "*Shout*" emitted to another agent holding resources encourages the agent to take "*Move*", which is a communicative action.

Note that it is not natural or easy for the agents to acquire such a peaceful policy since expressing the sense of a danger is accompanied by paying a cost. For example, in a competitive situation, if an opponent agent has not acquired a policy to avoid a danger, a fight may occur even when the agent produces a loud sound like the big sound representing a danger. At this time, the opponent agent cannot distinguish resources occupied by the agent from those existing freely in the environment, which will not make the opponent agent to take a "*Move*" action. A risk of the agent is increased in this situation, therefore, the policy "*Shout*" is difficult to be acquired.

3 Simulation Results

We conducted simulation studies. Our setup includes a few symbols each of which has no meaning before learning. Table 1 presents parameters and their values used in simulations.

Table 1: Parameters

Number of agents	100
Transition probability $P_i(i = 1...4)$ with the danger (none, danger, agent, resource)	(0.55,0.05,0.2,0.2)
without the danger (none, danger, agent, resource)	(0.6,0,0.2,0.2)
Time step per generation, t_{max}	100
Generation step, T_{max}	500
Sensor dimensionality, S_t	4
Action simensionality, A_t	4
Cost for making a sound, c_v	0.1
Damage by a danger, d_d	20
Damage by a fight, d	5
Resource in environment, R	5
Occupied resource, R_i	consumption 1 per a time step
Learning rate, α	0.9
Discounted rate, γ	0.9
Step size parameter, β	0.5
Number of breeding(dying) per generation	5
Gaussian noise, ϵ	N(0, 0.01)

Figure 2 shows the time course of fitness summed over all agents in each generation. Figure 3 shows histograms of actions selected at states during learning in an environment in which dangers exist, while Figure 4 shows histograms of actions selected at states during learning without dangers. As shown in Figure 3, communication emerged so that useless fighting that makes each agent's fitness to decrease were avoided. Hence, the sum of acquired fitness in each generation was larger in the environment with dangers than in that without dangers. Without dangers, we observed that the agents either ran away from their opponent or fought with it depending on their states including the

past experiences.

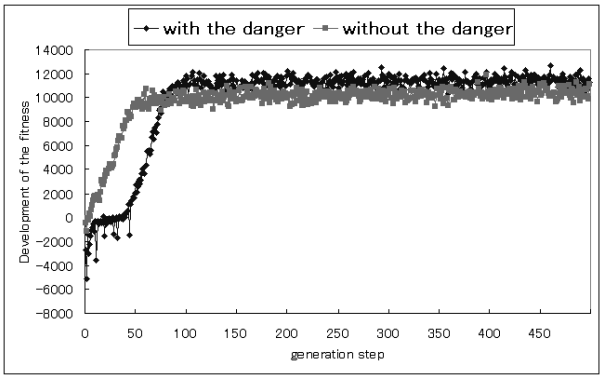

Figure 2: Development of the fitness

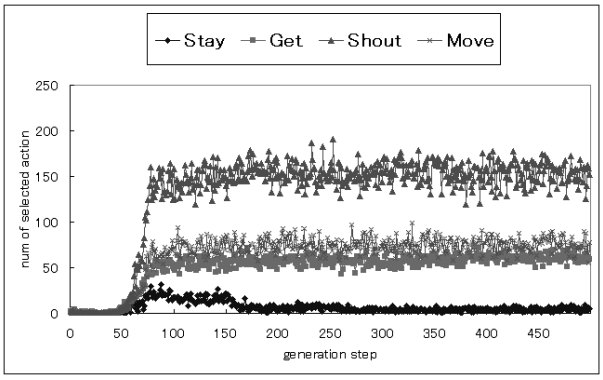

Figure 3: Development of actions by agents in an environment with dangers

4 Conclusion

We have demonstrated that the existence of dangers in the environment encourages the agents to develop communication appropriate for acquiring higher fitness by avoiding useless fights.

In a scientific view, it is believed that every organism have evolved the way "to use what can be used" in the history of living. All the symbols might be meaningless originally, but gradual assignment of meanings profitable for adapting to the environment makes the agents to utilize them; this would be the emergence of communication. The result shown in this article can be regarded as an example of such emergence process of communication.

In the future, we will try to examine the essence of communication such to increase communicative symbols of agents.

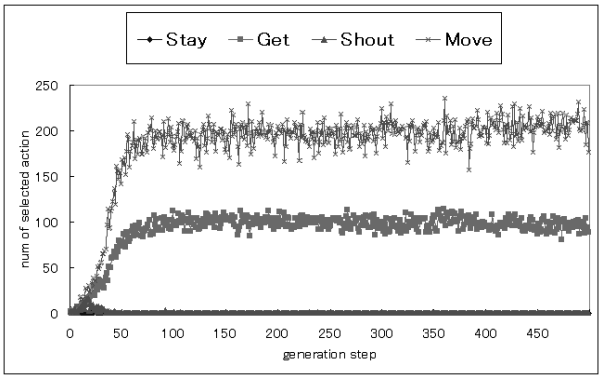

Figure 4: Development of actions by agents in an environment without dangers

References

[1] B.MacLennan: Synthetic Ethology -An Approach to the Study of Communication-, Artificial Life II, pp.631-658, 1991

[2] D.Ackley and M.Littman: Interaction Between Learning and Evolution, Artificial Life II, pp.487-509, 1991

[3] R.S.Sutton and A.G.Barto: "Reinforcement Learning : An Introduction", MIT Press, Cambridge, MA, 1998

[4] T.Sasaki and M.Tokoro: Adaptation of Evolutionary Agents towards the Dynamic Environment, Computer Software, Vol.14, No.4, pp.33-46, 1997