UNDERSTANDING COMMUNICATIVE INTENTIONS USING SIMULATED ROLE-REVERSAL

M. KLEIN*

Center for Language and Speech Technology, Radboud University Nijmegen, Nijmgen, 6500 HD, Netherlands
**E-mail: M.Q.Klein@let.ru.nl*

Understanding the communicative intention of a speaker is the ultimate goal of language comprehension. Yet, there is very little computational work on this topic. In this chapter a general cognitive plausible model of how an addressee can understand communicative intentions is presented in mathematical detail. The key mechanism of the model is simulated role-reversal of the addressee with the speaker, i.e., the addressee puts himself in the state of the speaker and — using his own experience about plausible intentions — computes the most likely intention in the given context. To show the model's computational effectiveness, it was implemented in a multi-agent system. In this system agents learn about which states of the world are desirable using a neural network trained with reinforcement learning. The power of simulated role-reversal in understanding communicative intention was demonstrated by depriving the utterances of speakers of all content. Employing the outlined model, the agents nevertheless accomplished a remarkable understanding of intentions using context information alone.

Keywords: Understanding Intentions; Communicative Intentions; Non-Verbal Communication; Multi-Agent-Systems.

1. Introduction

When a baby cries, the information content transmitted in the acoustic signal is very low[a]. Nevertheless, a mother can usually understand what the baby desires. She can do so because she understands (i) the context of the cry (last meal, state of diapers, etc.), as well as of (ii) the normal desires of a baby (to be fed, to be dry, etc). While utterances with such a low information content are exceptional, it is generally the case for almost every utterance that the literally transmitted information is not sufficient to understand the communicative goal of a speaker, but *context* and *likely*

[a]i.e., although the individuals cries might be quite different, these differences do not systematically related to a difference in content (at least not in the early stages of development).

desires are required as additional key parameters. To understand the communicative goal of a speaker is not a minor side issue, but it is the overall purpose of every act of inter-human communication. At the very fundament of human communication lies the understanding that a speaker (or, as in the example above, a crying baby) has a certain intention and wants you to understand this intention.[1] And to understand this intention, the context (including the current state and history of the speaker, as far as it is known the the addressee), as well as our estimation of likely desires of the speaker are essential sources of information. Only an approach integrating these can be considered a good model of human communication. In fact, our good understanding of each other, despite the fact that our utterances are so imprecise and sparse in terms of content can only be explained within the framework of such an integrated approach. Embedding cognitive processes involved in communication and language in a more general framework of processes concerned with the understanding of intentions is considered essential,[2,3] but so far very little computational work uses such an approach.

To understand intentions in the way described above requires a number of cognitive abilities. First of all, a person must have the ability to attribute a desire to another person, even if this desire is different from desire the attributing person has himself. This ability has been coined *Theory of Mind*.[4] This term is generally considered to include the second precondition of the model outlined above: the ability to regard actions as caused by those attributed inner states. Given that these two conditions are fulfilled, we can ask the question of *how* it is possible for a person to compute the underlying desire of an action.

The contemporary philosophical literature distinguishes two contrasting approaches to solve this problem: *theory-theory* and *simulation theory*.[5–7] While theory-theory would describes this computations as a detached theoretical process, simulation theory postulates that we simulate the mental state of the observed person in our own cognitive system. In other words, we put ourselves in the shoes of the other person. This means, for example, that we could estimate an emotional state of a person by simulating the situation or context of that particular person.

One of the main computational advantages of simulation theory over theory theory is that the machinery used for understanding an action is more or less the same as the machinery used for selecting your own action. The model I will present in this chapter draws heavily on this advantage. All the components an agent uses to understand an intention are the same

as those the agent uses for the selection of his own goal-directed actions. What I will present is a first simple computational approach to model the understanding of communicative intentions taking into account the context *and* likely desires. To demonstrate how effective these two parameters can be used in understanding a communicative intention, I will use communication signals that are utterly empty in terms of content (comparable to the cries of a baby), with the only information transmitted being that an act of communication has been made. A multi-agent system is used in which agents receives a reward if they are in a certain class of states of the environment. Using reinforcement learning,[8] the agents learn which states of the environment are desirable. Agents can perform a set of non-verbal actions to get into these desired states. In certain cases a desired state cannot be produced by an action of the agent itself, but by the action of another agent. In these cases, an agent is allowed to produce a communication signal without any content. The desired state that the signaling agents is trying to accomplish is considered the communicative intentions. The agent decides which action (among non-verbal and the one verbal) to perform by means of a Markov decision process using a value function and a pre-programmed forward model as it was described in previous work.[9] Using their own experience - their knowledge about which states of the environment are desirable, as well as their full awareness of the current state of the speaker, the addressees computes the plausible intentions of the speaking agents by a form of role-reversal. After putting themselves into the state of the speaker, the addressed agents use their forward model to test which of the plausible intention of the speaking agent they can actually bring about (assuming that the speaker wants them to bring about a certain state). Of those role-reversed states that the addressee is able to bring about, it is the one with the highest value that is considered the communicative intention of the speaker.

2. Method

Value Function

In the simulation work presented in this chapter, a value function $V()$ maps complete states of the simulated environment to a value (equation 1).

$$V^{\pi}(s_t) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\} \tag{1}$$

The value is an estimations of how good it is for an agent to be in this particular state (i.e. how much he desires the state). Values are positive or negative real numbers. $V^{\pi}(s_t)$ is the estimation of the value of state s_t at (discrete) time step t under a *policy* π.[10] Here, π is a mapping from states s and actions a to the probability $\pi(s, a)$ of performing action a when in state s. $V^{\pi}(s_t)$ is defined in terms of the expected sum of discounted rewards r. The expected value is taken with respect to the Markov chain $\{s_{t+1}, s_{t+2}, ...\}$ where the probability of transition from state s_{t+k} to s_{t+k+1} is given by π. Future rewards are *discounted* by the discount factor γ. The higher the value of γ, the more importance is given to later rewards, i.e. the less they are discounted (see Ref. 10 for a more detailed explanation of the formula and the theory that goes with it).

The value function is implemented as a single-layer feed-forward neural network. To train this network we used *TD(0) reinforcement learning*.[8] In TD-learning, the so-called TD-error gives the distance from the correct prediction and the direction of the deviation. Thus, it can be used to change the weights of a neural network. The TD-error δ is computed by subtracting the current state value of state s_t $V(s_t)$ from the sum of the reward r_{t+1} and the value of the next state $V(s_{t+1})$ times the discount factor (equation 2). Given δ, the value of the state $V(s_t)$ is changed to $V(s_t) + \alpha\delta$, where α is the rate of change (equation 3).

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$

$$V(s_t) \leftarrow V(s_t) + \alpha\delta \tag{3}$$

Action Selection

The value function allows to determine the *most desired state* of every agent in every state: the desired state is the state with the highest value. However, not every state can be reached from every other state. In fact, apart from the context state s_t only those few states are accessible which can be produced from s_t through a single action in a single time step. Therefore, the value function only needs to compute the value of those states which can be reached from the current state. To compute which states are accessible, or, in other words, to select a (verbal or non-verbal) action the *consequence* of actions needs to be estimated. This is accomplished with another device - a so-called *forward* model.[11] Within motor control, forward models are used to predict sensory consequences from efference copies of issued motor commands.[12] In the model described in this paper, we use forward models for the selection actions in the following way: the outcome of all possible

actions in the present context is predicted with the forward model and then the action which produces the most desired effect is chosen. F predicts a subsequent state s^*_{t+1} based on a current state s_t (context) and a possible non-verbal or verbal action (utterance) u^*_t.

$$s^*_{t+1} = F(s_t, u^*_t) \tag{4}$$

Given the forward model F, utterances and actions are selected by means of a function $\arg\max_u$ which selects the verbal or non-verbal action that produces the most desirable state (equation 5).

$$u_t = \arg\max_u [c(s_t, u^*_t) + V(F(s_t, u^*_t)] \tag{5}$$

This function returns that one from all possible a^*_t's which, given the context s_t, is mapped by the forward model F into a state s for which the value function V returns the highest value. Since $\pi(s, u)$ can be determined on the basis of the function described in equation 5, we will, for the rest of this article, no longer talk about π, but only about the forward model and the value function.

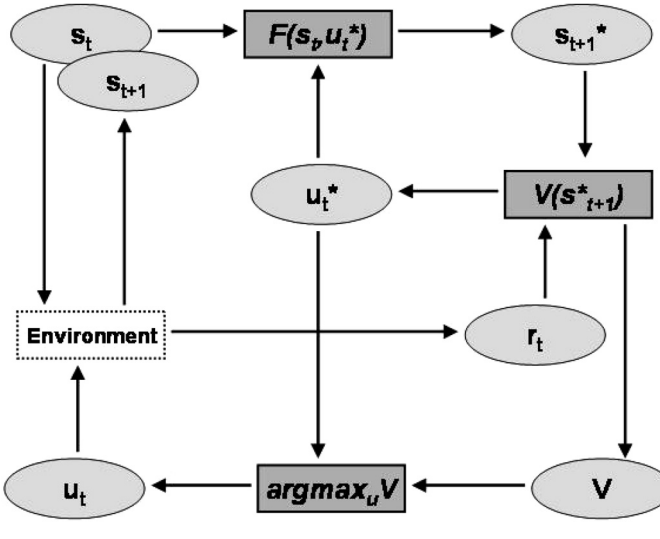

Fig. 1. The figure shows the architecture of action selection. In the current state s_t, the forward model $F()$ is used to predict the outcome of possible non-verbal or verbal actions. The value function $V()$ then estimates how desirable such an outcome is. The selected action is the one which leads to the most desirable outcome. After action selection, the environment determines the reward r and the next state s_{t+1}.

To be able to choose a *verbal* action, an agent needs to be able to compute the outcome of such a verbal action. In the simulations described in this chapter this is done in the following manner: Given the current state the speaker computes the outcome of all possible actions of a possible addressee with the pre-programmed value function and estimated the value of those outcomes with his trained value function. If any of the actions of a possible addressee leads to a state with a higher value than those states he can bring about himself he will choose to signal this addressee. This, of course, assumes (i) that the addressee will understand what the speaker want from him - which is only the case in later stages of training and (ii) that the addressee will actually cooperate. To keep things simple and we avoided all issues related to cooperation and made it a general policy of the addressee to cooperate.

Understanding Intentions

Here we state the mathematical and computational core of the theory presented in this paper. It is based in the following assumption:

(i) The addressee assumes (correctly in our simulations) that, if he is spoken to, the speaker desires that the addressee performs an action and that this action is the one that is the optimal action for the speaker in the current circumstances.

(ii) The value function of the addressee can serve as an approximation of the value function of the speaker, i.e. speaker and addressee desire similar things in similar situations.

Therefore, to understand the communication intention of a speaker, an addressee needs to (i) understand the current state of the speaker, including, of course, the speaker's environment. This is, of course, a highly idealized assumption. In the simulation presented in this chapter, agents, however, have full access to the complete state of the game. The state, however, needs to be role-reversed, i.e. the addressee needs to *put himself in the shoes* of the speaker. On the basis of this role-reversed current state, the addressee can find the action that is optimal for the speaker using his own value function to serve as an approximation of the value function of the speaker

$$a = argmax_{V_{sp}} a(F(s_c, a_{ad})) \tag{6}$$

The role-reversed value function is denoted by V_{sp}. I use the term *desire* and *intention* in the following manner. States of the world which the agents know to be beneficial for themselves are desired states, while states of the world which they are actually trying to reach by some action or

utterance are called intended states. In our theoretical framework an agent has many desires. However, only some of these desires actually become intentions. The desired state that triggered the verbal action is regarded as the communicative intention of the verbal action. If the addressee chooses an action that brings about this intended state he has correctly understood this intention.

The Acquisition Environment

We test our hypotheses about language acquisition and communication in a simulation of a multi-agent game. The goal in this game is to obtain food through verbal and non-verbal actions. In this simulation, *food* grows in certain intervals on *trees* (how this time interval is calculated is explained in the appendix). There are three trees $T_1 ... T_3$, growing three types of food. Every tree T_i can hold maximally 5 pieces of food. Time is supposed to advance in discrete jumps, from $t = 1$ to $t = 2$, $t = 2$ to $t = 3$ etc. Each two successive times t_i and t_{i+1} are separated by an action a_{t_i} of one of the agents, so that the state $s_{t_{i+1}}$ at t_{i+1} is the result that action a_{t_i} produces in the state s_{t_i}.

Within a certain time interval (d_o) invariably one piece of food gets *digested*, i.e. it disappears. Once the total amount of food in the game is below the threshold n_o, 3 pieces of food grow simultaneously on one of the three trees. Because of this design, the agents cannot afford to rest once they have gained a sufficient amount of food items. Agents never starve to death, but for every time step during which they do not have any food they get a very negative reward.

Agents can perform one of the following 12 actions:

- harvest a tree, i.e. collect all its food (3 possibilities)
- give one piece of food to another agent (2 other agents $\times$ 3 food types = 6 possibilities)
- send a communication signal to one of the other agents (2 other agents = 2 possibilities)
- *no action* (1 possibility)

At each transition between two successive times, only one agent can perform an action. This agent can perform either one non-verbal or one verbal action. Generally, the agents take turns. However, when an agent asks another agent for a type of food, the normal order of play is suspended for one time step and while the addressee gives (or fails to give) the desired

object to the speaker. An agent can only address one of the other agents, never both of them.

The goal of the agents in the game is to have at least one piece of each food type at all times. Therefore, the reward function was designed in the following way: Each agent gets a reward at every time step. If an agent has at least one item of every food type, he gets a reward of +3, otherwise he gets −1 for every food type which is missing in his store at that time.

3. Results

We performed a number of simulations during which the neural network based value function of the agents were trained and the percentage of correct understood communicative intentions were measured. Figure 2 shows a Hinton diagram of the weights of trained value function (at the end of the simulation). In that simulation a high γ-value was chosen and, as a result, the agents have learned that it is good to have more than one item of every type, although a direct reward is only given for the first item of each type. The diagram also shows that the agents all have a good understanding of

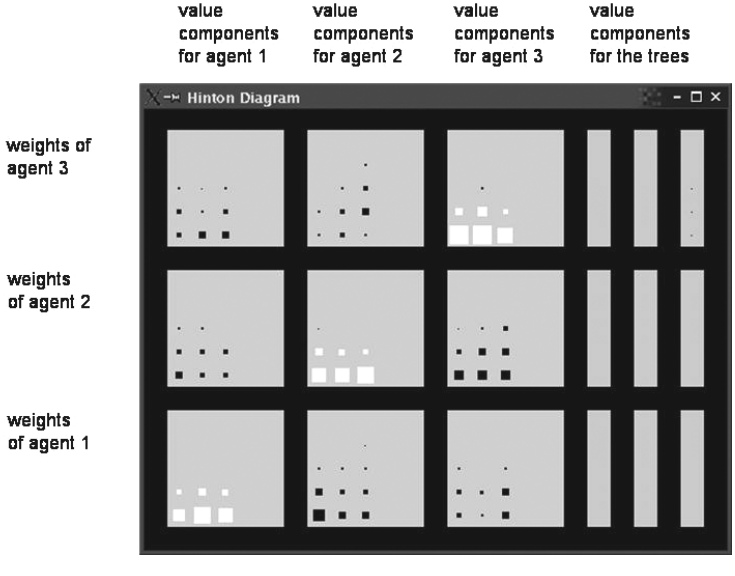

Fig. 2. This figure shows the weights of the value functions of the three agents for a γ - value of 0.9. The size of the squares represents the strength of the weights; the color represents the polarity (white is positive, black is negative).

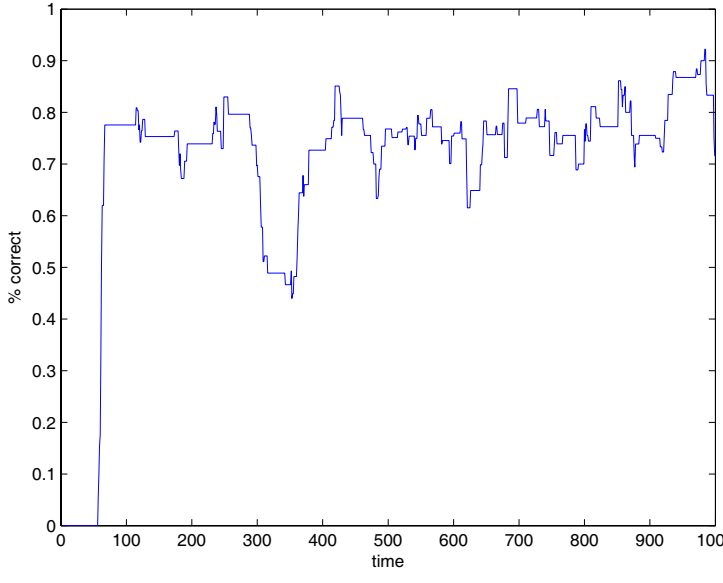

Fig. 3. The figure shows the average percentage of correctly understood communicative intentions over 15 runs.

which states of the game are desirable. Note, however, that there are subtle differences between the weights of each agents — even when the states are role-reversed the computed value will not be exactly the same.

This is also the reason why the number of correctly understood communicative intentions does not go up to 100%, but reaches a plateaux of about 80% after an initial fast increase of performance in the beginning (see Figure 3. This slight difference in value function is probably due to the fact that the weights are initialized randomly and for exploratory purposes during action selection a random number is added to the value of every action outcome. Nevertheless, given that no verbal information is given to the agents, the number of correctly understood utterances after a short training interval is remarkably high.

To illustrate the exact way the system works, two example *conversations* are shown here (see Figure 4 for the exact situations in which the two interactions took place). The first one is an incorrect case from early training (time step 2066 of 20000), i.e. the addressee does not understand the communicative intention of the speaker, due to his incompletely trained value function.

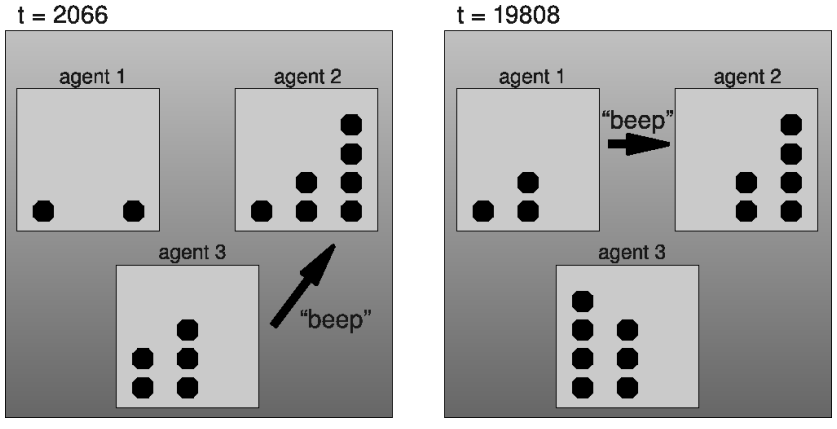

Fig. 4. This figure shows the situations of the two example interactions.

(i) Agent 3 needs food type 3. He correctly addresses agent 2 who is the only agent who has this type of food.
(ii) Agent 2 has items of all three food types. For each of the three food types he computes the consequent state should he give agent 3 an item of this type. Then, using role-reversal, he computes what value the three consequent states would have for agent 3. Due to his insufficient training he computes 0.2649376 for food type 1, 0.25863677 for food type 2, and 0.2633224 for food type 3. As a result, he gives agent 3 and item of food type 1 — clearly the wrong interpretation of the speaker's intention.

The second example is a correct case from the later stages of training (time step 19808 of 20000) when the addressee correctly understands the intention of the speaker.

(i) Agent 1 needs food type 3. He beeps agent 2, since agent 3 does not have food type 3.
(ii) Agent 2 has food type 2 and 3. He applies his value function (role reversed) to the outcome of the possible actions of giving agent 2 food type 2 (value: 1.2755736) or food type 3 (value 1.28508). Consequently, agent 2 gives food type 3 to agent 1 — the correct interpretation of the speaker's intention.

4. Discussion

This chapter introduced a general cognitive plausible theory of intention understanding in mathematical detail. Its effectiveness was demonstrated in a number of simulations using multi-agent systems. The estimation of intentions was performed with a value function implemented as a neural network and trained with reinforcement learning. To demonstrate the power of the approach we used utterances without content, so the only information an addressee did receive was that an utterance has been made. Nevertheless the amount of correctly recognized communicative intentions was around 80% after training.

One of the reasons for the recognition rate to be that high is the current implementation uses two major simplifications of the simulated world in comparison to real communication situations. The first one is that the state of the speaker and its context is fully accessibility to the addressee. The second one is that there is a close similarity between the value function of all agents. The similarity is accomplished by the fact that they are given exactly the same rewards and they also use the same γ parameter (i.e., they have the same attitude towards the relation between short term and long term goals). And while it can be generally assumed that all humans have somewhat similar goals just by the fact that they are the same species, difference in goals are given by genes and environment.

Simulations that do not use these simplifications are bound to be interesting and would be a possible extension of this work. However, when the value function of the agents start to differ due to differences in experience and hard-wired parameters, agents need to rely stronger on the verbal content of an utterance to determine the communicative intention. Therefore, a model needs to be developed that can use information given literally in an utterance (as in previous work[9]) together with the context and information obtained through role-reversal.

References

1. H. P. Grice, *Philosophical Review*, 377 (1957).
2. M. Tomasello, *Constructing a Language - A Usage-Based Theory of Language Acquisition* (Harvard University Press, 2003).
3. S. C. Levinson, On the human interaction engine, in *Roots of Human Society*, eds. N. J. Enfield and S. C. Levinson (Berg, 2006).
4. D. G. Premack and G. Woodruff, *Behavioral and Brain Sciences*, 1, 515-526 (1978).
5. A. Goldman, *Behavioral and Brain Sciences*, 16: 15-28 (1993).

6. J. Decety and P. L. Jackson, *Behavioral and Cognitive Neuroscience Reviews*, 3, 71-100 (2004).
7. S. D. Preston and F. B. M. de Wall, *Behavioral and Brain Sciences*, 25, 1-72 (2002).
8. R. S. Sutton, *Machine Learning* **3**, 9 (1988).
9. M. Klein, H. Kamp, G. Palm and K. Doya, *Neural Networks* (2008), submitted.
10. R. S. Sutton and A. G. Barto, *Reinforcement Learning - An Introduction* (MIT Press, 1998).
11. M. Jordan and D. E. Rumelhart, *Cognitive Science* **16**, 307 (1992).
12. M. Kawato, *Current Opinion in Neurobiology*, 718 (1999).