Expressing and Understanding Desires in Language Games

Michael Klein[1,2], Hans Kamp[2], Guenther Palm[3], and Kenji Doya[1,4]
[1]ATR Computational Neuroscience Laboratories
[2]Institute for Natural Language Processing, Stuttgart University
[3]Department of Neural Information Processing, Ulm University
[4]Crest, Japan Science and Technology Agency

Abstract

We speak because we want to get certain things accomplished. In this study we present the simulation of a multiagent language game which takes this into account. In contrast to previous language game studies, our agents use reinforcement learning to learn a function assigning a value to every state of the game. This value, that tells the agent how desirable the state is, is used along with a forward model to select actions. The agent can select verbal and non-verbal actions, depending on whether speaking or manipulating the world *directly* is more likely to bring about the change which the agent desires. On top of these capabilites, we used two rule-based agents to train a language learner. The learner trains a forward model of context-dependent utterance effects, which he then uses to express his desires and understand the desires of other players.

Introduction

Verbal communication normally serves some purpose (Wittgenstein, 1953; Austin, 1961). We speak because we want to get certain things accomplished. And often our concern is to get our addressees to do something for us, or to assist or counsel us in connection with our own actions. It is in such a setting of purposeful language use that language is acquired. As soon as a child begins to make use of language, it does so, most of the time, in the hope of getting its environment to comply with its wants and desires. And when it finds out that langauge can serve this purpose quite well, it will be all the more motivated to improve its linguistic skills further. In particular, this will impel it to learn the linguistic conventions that determine which expressions correspond to which states of affairs - conventions that are observed by those with whom the child interacts and whom provide the inputs to its acquisition process. Once these conventions have been mastered, the child can choose the utterances that express the states of affairs it desires in accordance with these conventions, with the effect that the one or ones it addresses will tend to understand what it wants and do what is needed to bring the wanted state about. Therefore, it seems reasonable enough to assume that the interplay between what linguistic expressions *mean* and the purposes to which language is put is an important, and probably crucial factor to the way in which human languages are learned. So it seems natural to try and see if the essential features of such a learning situation, in which the correlation between the meaning of language and the extralinguistic purposes of its use is of paramount importance, can be modelled in a somehow testable manner. Such a model should place the language-acquiring agent together with other agents who already possess the linguistic skills he still has to acquire in a common environment, and should explore the conditions that are needed for the agent to acquire the lingusitic knowledge through interaction with the agents which have mastered them already.

Surely, however, linguistic conventions are not the only things that children learn. They must also learn what states are desirable. Not just because they are agreeable in themselves, but also because of the subsequent rewards they promise.

In this study we build a computational model of language acquisition, in which two agents with a rule-based language module train one language learner. Using supervised learning, the learner trains a *forward model* which predicts the (context-dependend) effects of utterances. To train this model the learner observes the communication of the other agents. He can then use this model (i) to find the right utterance to express his own desires, and also (ii) to understand other agents by mapping speakers' utterances on the state they desire. We used a game environment, because it allows the agents to form their own desires. Agents learned a *value function* that can assign a value to every states telling the agent how desirable it is. Along with the forward model this function is used to select actions. It gives the agent the freedom to speak or not speak. In other words, the agent needs to decide for himself whether speaking or manipulating the world *directly* is the best action in a situation. This and the *effect-oriented* approach to language are the main differences between our study and previous studies in language games (Steels, 1996; Steels, 2001).

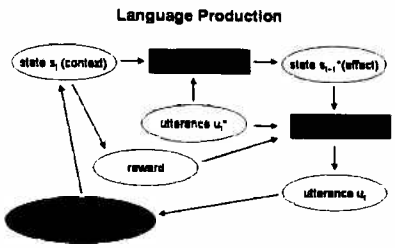

Figure 1: A simplified graph of how utterances (and other actions) are selected in our model: The forward model predicts the context-depended utterance effects and the value function determines whether the utterance is selected on the basis of how desirable the effect is.

Theoretical Framework

In order that an agent can express desires[1], he must have desires that he can express. That is, the agents of our model need to have the possibility of representing desires, and that in a form that is independent of the language that they use to express them and that the learning agent is to acquire. The form we have chosen is that of representing desires as *valued* states of the world. For every state of the world, the agent has a positive or negative numerical value expressing how much it desires this state. A function mapping every state on such a value is called a value - function (equation 2). Such a function is an estimation of how good it is for an agent to be in a given state. The notion of how good is defined in terms of future rewards that can be expected, or, to be precise in terms of the *expected return* (Sutton and Barto, 1998). The expected return is the sum of discounted rewards, which can be expected in and after a certain state (equation 1) The γ-parameter is the discount factor, which determines the value of future rewards. Given this definition of the expected return, the value function is defined in equation 2.

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{1}$$

$$V^{\pi}(s) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\} \tag{2}$$

Reinforcement learning is a suitable method of generating such a function during interaction with the environment.

These methods allow us to determine the *desired state* of every agent in every state: It is the state with the highest value. But not every state can be reached from a particular other state. Therefore, we need a *desire hierarchy* with more desired and less desired states. The value function gives us exactly what we need.

We propose a theory of communication and language acquisition, with the the following five essential points:

(i) An agent experiences utterances used by others speakers to have effects on the observable world. These effects of utterances are dependent on the context.

(ii) These effects on the observable world are indirect and are achieved by a more direct effect on unobservable states, such as the mental states of the addressees.

(iii) Linguistics events, i.e. experienced context-dependent effects of utterances are used to train a function mapping context configurations and utterances to effects.

Such a function, which predicts a sensation s^*_{t+1} based on the state s_t (context) and the action u_t (utterance) has been be called a *forward model* (Jordan and Rummelhart, 1992).

$$s^*_{t+1} = F(s_t, u^*_t) \tag{3}$$

(iv) A speaker uses a certain expression, because he desires the effects he expects the expression to produce in the present context (according to his experience). By using his forward model and his value function, he chooses the action which will lead to the state of the world which he desires most.

This output function (or utterance function) maps the observed state and the desired observation into an utterance (equation 4).

$$u_t = argmax_u V(F(s_t, u^*_t)) \tag{4}$$

(v) An addressee understands an utterance by using his forward model applied to the context and the utterance with which he was addressed. He thereby maps the utterance on the expressed desire or the intention of the speaker. (This is based on the plausible assumption, that human speakers share a core model of context-depended utterance effects.)

In other words, an addressee understands what the speaker is trying to achieve by using his *predictor*.

From this theoretical framework, we derive the following two hypotheses: (i) In a game environment, where only certain accomplishments are rewarded, agents equipped with value function (trained with reinforcement learning), a (rule-based) forward model, and a set of verbal and non-verbal actions can learn to behave in an optimal way, employing language and other actions whenever appropriate. (ii) In such an environment, an agent equipped with a optimal value

[1] We use the term *desire*, instead of *goal*, as we regard *desire* as the broader term. Desires can sometime have the form of a goal state, but in general can be a set of states instead of a goal point.

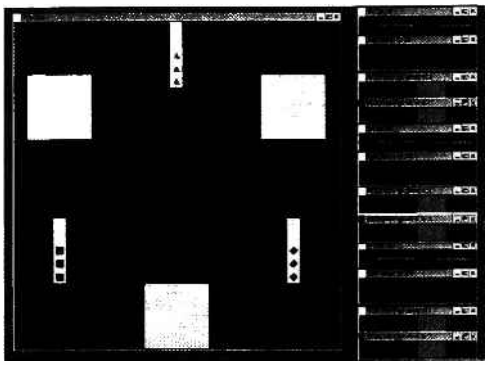

Figure 2: This shows the initial game state. The long yellow standing rectangles are the trees, each holding 3 pieces of food. The grey squares are the agents. They have the capacity of storing 5 pieces of each food type. The bar on the right displays scores and utterances. The green bars show, which agents cooperated with which other agents in their last move.

function and a rule-based model for non-verbal actions, can learn to use language to achieve his goals by expressing his desires and to understand the desires of other agents.

The Game

We test our hypotheses about language acquisition and communication in a multi-agent simulation. In this simulation, *food* grows in certain intervals in *trees*. In the present work, we use three trees growing three types of food. Every tree can hold maximally 5 pieces of food, and 3 pieces of food grow simultaneously, once the amount of food in the game is below a certain threshold.

There are three agents in the game. Every agent can store 5 pieces of each food type. Always after a certain time interval one piece of food gets *digested*, i.e., it disappears. This is to guarantee, that the agents need to act and cannot rest, after they have gained a sufficient amount of food items. However, they do not *starve* if they have no food for a number of time steps, but they get a low reward.

Agents can perform one of the following actions:

- harvest tree (take down all the food)
- give one piece of food to another agent
- ask another agent for a type of food
- do nothing

Generally, the agents take turns. However, when an agent asks another agent for a type of food, the normal order pauses for one time step, as then it is the turn of the addressee to give (or not to give) the desired object to the speaker. In our simulation, agents give objects away, when they are asked to. We do not address the emergence of cooperation, but assume a cooperative attitude from the very start. Of course, we agree, that children must learn to become competent social agents, e.g. to abide by the social conventions of certain forms of give and take, and to understand how the use of language to get others to do things for you fits in with social interactions generally. However, a small, language learning child is not expected to be on equal footing with those from whom it learns, and certainly is not expected to give as much as it gets. In fact, children would have considerable difficulties to master language if e.g. their parents would not behave in a beneficial way.

The goal of the agents in the games is to have one piece of each food types at every time step. 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, it gets a reward of $+3$, otherwise it gets -1 for every food type which is missing in his store at that time step.

The agents in the game are simulated independent of each other. Every agent observes the relevant features of the environment at every time step. To detect state changes correlating with utterances, an agent needs to memorize past states and utterances. Therefore, every agent has his own short term memory, storing the complete observable game state (including the utterances) for a constant number m of time steps. Further, every agents is equipped with long term memory in the form of weights between neural units storing the value function and the forward model. The agents interact with the world and other agents only by their perception, actions and utterances.

An utterance of an agent is defined by its content (i.e. which word is used), its speaker (the agent), and its addressee. An agent can only address one of the other agents, never both of them, but the third party can observe every utterance that is made. This important, because in our setting language is learned by observing the context-dependent effects of the utterances of other agents. Who an agent talks to and what he says, is up to him. The content of an utterance is defined by the *vocabulary* of the agents in the game. In the present study, it consists of the three words *triangle*, *square*, and *diamond*. The content of an utterance can only be one word. An agent can use only one utterance at every time step.

With respect to their linguistic capabilities, agents can either be a *teacher* or a *learner*. Teacher-agents use a rule-based dialogue system to produce and understand utterances. Teachers in this game have no interest in the learner-agents learning of language. They do not perform any actions or utterance with the goal to teach language to the learners. Instead, learner-agents can learn language by observing the language use of the teacher-agents.

To chose their actions (or utterances), the agents predict the outcome of the action in the present context with a forward-model. The forward model is rule-based for *no action*, harvesting trees, donating objects, and for the verbal

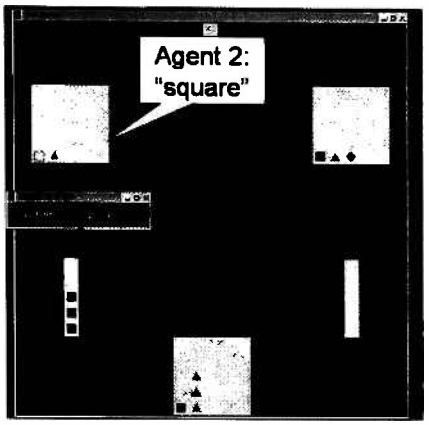

Figure 3: This shows an arbitrary state during the early stages of training. The last action of agent 1 was to ask agent 2 for the square. Obviously, this is not the best move. A better move would be to harvest the *square tree*, as with this action, the agent would get 3 squares instead of one. The agent has not correctly estimated the values of these two states and, therefore, has chosen the wrong action.

actions of the teacher-agents.

These outcomes are evaluated with the value function and the action (verbal or non-verbal) which will bring about the state with the highest value is chosen.

If a verbal action is selected and the addressed agent is a teacher, then the addressed agent will give the desired object to the speaker. If the addressed agent is a language learner, this agent applies its *forward model* to the utterance and the game state. With this model, he can estimate what kind of change the speaker desires, i.e. he computes the intention from the utterances and the context. In other words, the language learner understands the utterance, because he *wonders* what effect, according to its own experience, such an utterance has in the present context. Using the present state of the game and the estimation of the desired state of the speaker, the addressee then uses a rule-based algorithm to computed which action would bring this desired state of the game about.

Learning Algorithms

The value function maps states of the game to real numbers. A state s is a 6-tuple of three $m \times n$ binary matrices (one for each agent) and three n-dimensional binary vectors (one for each tree). In all simulations described in this paper $m = 3$ and $n = 5$. At every point in time t, every agent can perceive the complete state $s(t)$.

$$s = < A_1, A_2, A_3, T_1, T_2, T_3 > \quad (5)$$

$$A_i = \mathbf{A}_{mn} \quad (6)$$

$$T_i = \mathbf{t}_n \quad (7)$$

The value function is implemented as a neural network with one neuron for every binary value of the vectors and the matrices of s. The output of the network is the linear combination of the weighted binary inputs. To train the value function, we used TD(0) reinforcement learning (Sutton, 1988) as described in equation 8. The term given in 9 is the so-called TD-error, giving distance and direction to the correct prediction and determine the weight changes.

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (8)$$

$$r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (9)$$

Because of the huge number of possible states, we used a neural network function approximation of the value function. As exploration mechanism we used a *softmax* method.

The linguistic capabilities of the language learner in the game are represented by a forward model. This model learns the context-dependent consequences of utterances. The context of each utterance is the game-state, as described above. The forward-model is implemented by a single-layer perceptron, mapping utterances and game-states onto game-states. We used supervised learning to train this forward model (equations 10, 11, and 12).

$$e_k = y_k^* - y_k \quad (10)$$

$$\delta w_{ik} = \alpha e_k x_i \quad (11)$$

$$w_{ik} \leftarrow w_{ik} + \delta w_{ik} \quad (12)$$

Note that the same forward model can be trained by observing the effects of other agent's utterance as well as the effect of the agents own utterances (i.e. in our approach these two types of predictions are not distinguished, which of course is a considerable simplification).

Results

In the first stage of training, only the value function is trained. The value function enables the agents to estimate which states are desirable. At this stage, agents use a hard-coded forward-model to compute the outcome of possible action. They chose the action which brings about the outcome with the highest value. At the very early stages of the training, agents very often selected *no action* or senseless actions (such as donating objects to other players without being asked). Sensible, but suboptimal actions, as described in figure 3, did occur in the intermediate stage of training. After training, no more suboptimal action could be detected. Agents performed extremely well (approximately at the level of a human player or even better). Agents used language if appropriate, harvested trees whenever possible and optimal. They did not choose no action any longer, as usually some action or request would improve the state of agent. The value function gave the appropriate desire-[illegible]. We tested the performance of the model with

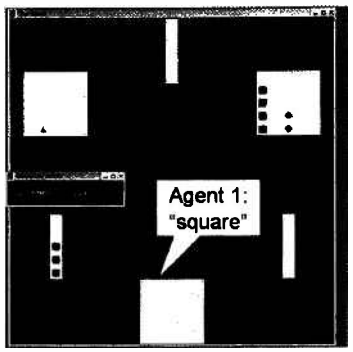

Figure 5: This shows an episode during language learning. The language learner (agent 2) asks agent 1 for the square, although agent 1 does not have one. Although the agent has a fully trained value function telling him which states are desirable, the forward model is not yet capable of predicting effects of the agent's utterances correctly. Therefore, the language learner is not able to understand the the context conditions and the normal *effects* his utterance has.

gamma = 0.9 / 0.7 / 0.5 / 0.3 / 0.1 (figure 4). The model showed stable performance for all values. The TD-error decreased faster with smaller *gamma*, while the overall performance was better with larger gamma. Closed to optimal performance could already be observed after about a million time steps.

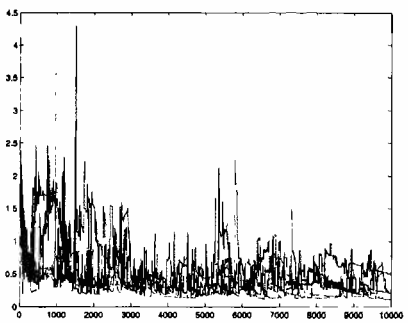

Figure 4: This is the development of the TD-error with $\gamma = 0.3$. The figure shows the development for 5 runs and 10^7 time steps (x is time in ksteps)

In the second stage of training, the forward-model of a language learning agent was trained. He was placed in an environment with two *teachers*. The initial language capabilities of the learner could best be described by *random utterances* and no understanding. Random utterances means that although the learner is a fully trained game player (i.e. he knows which state should be achieved), he has no idea which utterances are likely to bring these states about. This leads to situations, where e.g. he wants an object of one

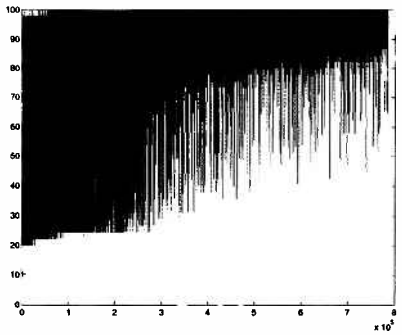

Figure 6: The prediction error of language learning changes over time. This graph shows the percentage of correct prediction through approximately 5 * 10^6 learning episodes

agent, but addresses the other agent, asking for a different object. Figure 5 illustrates another example where the language learner asks an agent for an object, which the addressed agent does not even have. This is because initially and in the early stages of training the learner has no or no accurate representation of the context condition for successful use of an utterance and also does not know its conventional effect. The language understanding of the learner faces the same problem. In early stages of training he usually shows no reaction upon being asked for a food item, because he wrongly maps the utterance to desired states of the speaker, that are unachievable from the current state or, in rare cases, he shows wrong reactions. While the training progresses, however, reactions are more and more in accordance with the utterances he is addressed with. Also the learner's utterances are more and more suitable with the situation and the desire until after training no difference between teacher and learner can be detected. As can be seen figure 6, the prediction error decreased very fast to a level close to 0.

Discussion

From the theoretical framework of human communication and language learning sketched above, we derived and tested two hypotheses: (i) In a game environment, where only certain accomplishments are rewarded, agents equipped with value function (trained with reinforcement learning), a (rule-based) forward model, and a set of verbal and non-verbal actions can learn to behave in an optimal way, employing language and other actions whenever appropriate. (ii) In such an environment, an agent equipped with a optimal value function and a rule-based model for non-verbal actions, can learn to use language to achieve his goals by expressing his desires and to understand the desires of other agents. We were able to confirm both hypotheses in our setting.

The aim of this simulation was to test our ideas in an environment which resembles the environment of humans. Of course, our simulation can only be a first approximation. A

considerable simplification was our decision to enable our agents to observe the complete environment at all times (including the object possession of other agents). However, such a setting is justified and necessary at this stage, as in early language acquisition the effects of utterances need to be observable by the learning children.

In general, it remains difficult to compare the behavior and architecture of our agents with the human neural system and human verbal behavior. While the computational model of reinforcement learning we used is generally considered an appropriate model for human learning, so far no behavioral human data which is directly comparably to our task has been obtained. Although our system generally allows a human subject to play the game, it is not suitable for children in the early stages of language development, and it is clearly in these early stages, where children learn the nature of language use. Even if such a detailed comparison between the behavior of the model and human infants is not possible at present, our theoretical ideas are well in accord with the ideas of psycholinguists on language acquisition (Tomasello, 2000) and we look forward to give a more detailed comparison, when suitable behavioral data has been obtained.

Concerning the relation of our neural architecture to the real human brain we have refrained from claiming to model certain brain regions. However, there is considerable evidence that the basal ganglia is involved in reinforcement learning, while the cerebellum performs supervised learning and acts as a forward model in certain tasks (Doya, 1999). These two learning types are the ones involved in our computational model. Of course, it would be wrong to assume, that these two brain structures handle language acquisition alone, but they need to interact with the cerebral cortex. The cortex would be involved in tasks we did not model in this study, but which are nevertheless at the core of language. The unsupervised learning of the cortex plays a major role in the acquisition of word forms and their association with concepts (Klein and Billard, 2001) and is necessary to form concepts in the first place (Klein and Kamp, 2002). Word forms have been taken as given in this study, while the idea of concepts has been totally ignored. However, in real language acquisition concepts play a important role and an extension of our model able to generalize and learn more abstract feature of the predicted effects of utterances is very likely to make language learning faster and more efficient. The exact relation between the structures of our model and real brain structures has to be worked out in more details. When we have better hypothesis, we would like to evaluate them with functional brain imaging methods.

Finally, we would like to mention, that, although the language in our present study is restricted to single-word requests, the framework is can be extended to multi-word utterances and especially different kinds of speech acts, such as questions and the sharing of information. The next step in our research is to test this considerably more complex linguistic challenge.

Acknowledgments

This work was supported by the German Academic Exchange Service (DAAD), the German Research Foundation (DFG), and the National Institute of Information and Communications Technology of Japan (NICT).

References

Austin, J. L. (1961). *Philosophical Papers*. Oxford University Press.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, 12:961–974.

Jordan, M. and Rummelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

Klein, M. and Billard, A. (2001). Words in the cerebral cortex - predicting fmri-data. In *Proceedings of the 8th Joint symposium on neural computation - The brain as a dynamical system, San Diego*.

Klein, M. and Kamp, H. (2002). Individuals and predication - a neurosemantic perspective. In Katz, G., Reinhard, S., and Reuter, P., editors, *Sinn und Bedeutung 6, Proceedings of the sixth meeting of the Gesellschaft fuer Semantik, Osnabrueck, Germany, October 2001*.

Steels, L. (1996). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multi Agent Systems*, pages 338–344. AAAI Press.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent systems*, pages 16–22.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning - An Introduction*. MIT Press.

Tomasello, M. (2000). First steps towards a usage-based theory of language acquisition. *Cognitive Linguistics*, 11:61–82.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.