5

Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations

Naoto Iwahashi
*National Institute of Information and Communications Technology,
Advanced Telecommunications Research Institute International
Japan*

1. Introduction

Recent progress in sensor technologies and in an infrastructure for ubiquitous computing has enabled robots to sense physical environments as well as the behaviour of users. In the near future, robots that change their behaviour in response to the situation in order to support human activities in everyday life will be increasingly common, so they should feature personally situated multimodal interfaces. One of the essential features of such interfaces is the ability of the robot to share experiences with the user in the physical world.

This ability should be considered in terms of spoken language communication, which is one of the most natural interfaces. The process of human communication is based on certain beliefs shared by those communicating (Sperber & Wilson, 1995). Language is one such shared belief and is used to convey meaning based on its relevance to other shared beliefs. These shared beliefs are formed through interaction with the environment and other people, and the meaning of utterances is embedded in such shared experiences. From this viewpoint, spoken language interfaces are important not only because they enable hands-free interaction but also because of the nature of language, which inherently conveys meaning based on shared experiences. For people to take advantage of such interfaces, language processing methods must make it possible to reflect shared experiences.

However, existing language processing methods, which are characterized by fixed linguistic knowledge, do not make this possible (Allen et al., 2001). In these methods, information is represented and processed by symbols whose meaning has been predefined by the machines' developers. In most cases, the meaning of each symbol is defined by its relationship to other symbols and is not connected to perception or to the physical world. The precise nature of experiences shared by a user and a machine, however, depends on the situation. Because it is impossible to prepare symbols for all possible situations in advance, machines cannot appropriately express and interpret experiences in dynamically changing situations. As a result, users and machines fail to interact in a way that accurately reflects shared experiences.

To overcome this problem and achieve natural linguistic communication between humans and machines, we should use methods that satisfy the following requirements.

Source: Human-Robot Interaction, Book edited by Nilanjan Sarkar,
ISBN 978-3-902613-13-4, pp.522, September 2007, Itech Education and Publishing, Vienna, Austria

Grounding: Beliefs that machines have and the relationship among the beliefs must be grounded in the experiences and the environment of each user. The information of language, perception, and actions should be processed in an integrative fashion. The theoretical framework for language grounding was presented by Roy (Roy, 2005). Previous computational studies explored the grounding of the meanings of utterances in conversations in the physical world (Winograd, 1972; Shapiro et al., 2000), but they did not pursue the learning of new grounded concepts.

Scalability: The machines themselves must be able to learn new concepts and form new beliefs that reflect their experiences. These beliefs should then be embedded in the machines' adaptively changing belief systems.

Sharing: Because utterances are interpreted based on the shared beliefs assumed by a listener in a given situation, the shared beliefs assumed by a user and a machine should ideally be as identical or consistent with each other as possible. The machine should possess a mechanism that enables the user and the machine to infer the state of each other's belief systems in a natural way by coordinating their utterances and actions. Theoretical research (Clark, 1996) and computational modelling (Traum, 1994) focused on the sharing of utterance meanings among participants in communication and have attempted to represent it as a procedure- and rule-driven process. However, we should focus on the shared beliefs to be used in the process of generating and understanding utterances in a physical environment. To achieve robust communication, it is also important to represent the formation of shared belief systems with a mathematical model.

All of these requirements show that learning ability is essential in communications. The cognitive activities related to the grounding, scalability, and sharing of beliefs can be observed clearly in the process of language acquisition by infants as well as in everyday conversation by adults. We have been developing a method that enables robots to learn linguistic communication capabilities from scratch through verbal and nonverbal interaction with users (Iwahashi, 2003a; Iwahashi, 2003b; Iwahashi, 2004), instead of directly pursuing language processing.

Language acquisition by machines has been attracting interest in various research fields (Brents, 1996), and several pioneering studies have developed algorithms based on inductive learning using sets of pairs, where each pair consists of a word sequence and non-linguistic information about its meaning. In several studies (Dyer & Nenov, 1993; Nakagawa & Masukata, 1995; Regier, 1997; Roy & Pentland, 2002; Steels & Kaplan, 2001), visual rather than symbolic information was given as non-linguistic information. Spoken-word acquisition algorithms based on the unsupervised clustering of speech tokens have also been described (Gorin et al., 1994; Nakagawa & Masukata, 1995; Roy & Pentland, 2002). Steels examined (Steels, 2003) the socially interactive process of evolving grounded linguistic knowledge shared by communication agents from the viewpoint of game theory and a complex adaptive system.

In contrast, the method described in this chapter focuses on fast online learning of personally situated language use through verbal and nonverbal interaction in the real world. The learning method applies information from raw speech and visual observations and behavioural reinforcement, which is integrated in a probabilistic framework. Through verbal and nonverbal interaction with a user, a robot learns incrementally and online speech units, lexicon (including words referring to objects and words referring to motions of moving objects), grammar, and a pragmatic capability. A belief system including these beliefs is represented by a dynamic graphical model (e.g., (Jordan & Sejnowski, 2001)) that has a structure that reflects the state of the user's belief system; thus, the learning makes it possible for the user and the robot to infer the state of each other’s belief systems. The method enables the robot to understand even fragmentary and ambiguous utterances of users, act upon them, and generate utterances appropriate for the given situation. In addition, the method enables the robot to learn these things with relatively little interaction. This is also an important feature, because a typical user will not tolerate extended interaction with a robot that cannot communicate and because situations in actual everyday conversation change continuously.

This chapter is organized as follows. Section 2 describes the setting in which the robot learns linguistic communication. Section 3 describes the methods of extracting features from raw speech and image signals, which are used in learning processes. Section 4 explains the method by which the robot learns speech units like phonemes or syllables. Section 5 explains the method by which the robot learns words referring to objects, and their motions. Section 6 explains the method for learning grammar. Section 7 addresses the method for learning pragmatic capability. Section 8 discusses findings and plans for future work.

2. Setting for Learning Interaction

The spoken-language acquisition task discussed in this work was set up as follows. A camera, a robot arm with a hand, and the robot’s head were placed next to a table. A user and the learning robot saw and moved the objects on the table as shown in Fig. 1. The head of the robot moved to indicate whether its gaze was directed at the user or at an object. The robot arm had seven degrees of freedom and the hand had four. A touch sensor was attached to the robot's hand. A close-talk microphone was used for speech input. The robot initially did not possess any concepts regarding the specific objects or the ways in which they could be moved nor any linguistic knowledge.

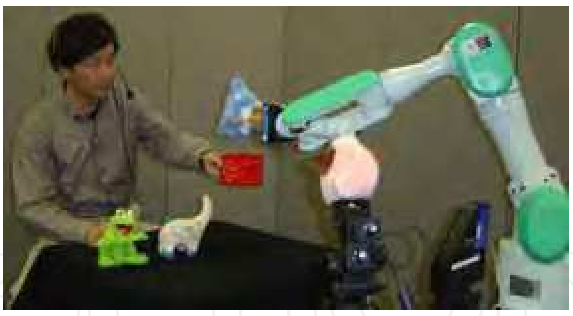


Figure 1. Interaction between a user and a robot

The interactions for learning were carried out as follows. First, to help the robot learn speech units, the user spoke for approximately one minute. Second, in learning image concepts of objects and words that refer to them, the user pointed to an object on the table while speaking a word describing it. A sequence of such learning episodes resulted in a set of pairs, each composed of the image of an object and the word describing it. The objects used included boxes, stuffed and wooden toys, and balls (examples are shown in Fig. 2). In each of the episodes for learning motions and words referring to them, the user moved an object while speaking a word describing the motion. Third, in each of the episodes for learning grammar, the user moved an object while uttering a sentence describing the action. By the end of this learning process, the user and the robot had shared certain linguistic beliefs consisting of a lexicon and a simple grammar, and the robot could understand utterances to some extent. Note that function words were not included in the lexicon. Finally, in the pragmatic capability learning process, the user asked the robot to move an object by making an utterance and a gesture, and the robot responded. If the robot responded incorrectly, the user slapped the robot's hand. The robot also asked the user to move an object, and the user acted in response. Through such interaction, the robot could learn the pragmatic capability incrementally and in an online manner.

These processes of learning lexicon, grammar, and pragmatic capability could be carried out alternately.

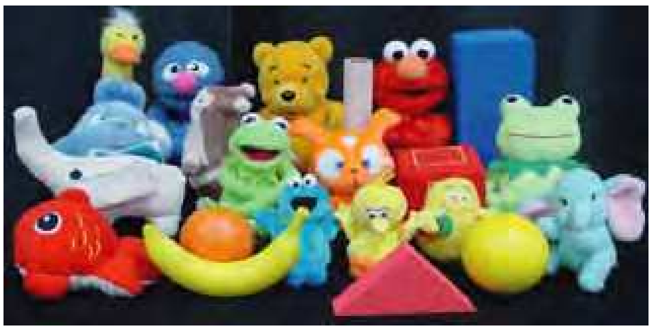


Figure 2. Examples of objects used

3. Speech and Image Signal Processing

All speech and visual sensory output was converted into predetermined features. Speech was detected and segmented based on changes in the short-time power of speech signals. The speech features used were Mel-frequency cepstral coefficients (Davis & Mermelstein, 1980), which are based on short-time spectrum analysis, their delta and acceleration parameters, and the delta of short-time log power. These features (25-dimensional) were calculated in 20-ms intervals with a 30-ms-wide window.

The camera contained three separate CCDs, enabling the robot to obtain three-dimensional information about each scene. The information regarding the object's position in terms of the depth coordinates was used in the attention-control process. Objects were detected when they were located at a distance of 50–80 cm from the camera. The visual features used were L*a*b* components (three dimensions) for the colour, complex Fourier coefficients (eight dimensions) of 2D contours for the shape (Persoon and Fu, 1977), and the area of an object (one dimension) for the size. The trajectories of objects were represented by time-sequence plots of their positions on the table (two-dimensional: horizontal and vertical coordinates), velocities (two-dimensional), accelerations (two-dimensional)..

4. Learning Speech Units

4.1 Difficulty

Speech is a time-continuous one-dimensional signal. The robot learns statistical models of speech units from such signals without being provided with transcriptions of phoneme sequences or boundaries between phonemes. The difficulty of learning speech units is ascribed to the difficulties involved in speech segmentation and the clustering of speech segments into speech units.

4.2 Method using hidden Markov models

It is possible to cope with the difficulty described above by using hidden Markov models (HMMs) and their learning algorithm, the Baum-Welch algorithm (Baum et al., 1970). HMMs are a particular form of dynamic graphical model that statistically represents dynamic characteristics of time-series data. The model consists of unobservable states, each of which has a probability distribution of observed data, and the probabilities of transitions between them. The Baum-Welch algorithm makes it possible to segment speech, cluster speech segments, and learn the HMM parameters simultaneously.

In this method, each speech-unit HMM includes three states and allows for left-to-right transitions. Twenty speech-unit HMMs were connected to one another to construct a whole speech-unit HMM (Fig. 3), in which transitions were allowed from the last states of the speech-unit HMMs to the first states. All parameters of these HMMs were learned using speech data approximately one minute in length without any phoneme transcriptions. After the speech-unit HMMs had been learned, the individual speech-unit HMMs h_1, h_2, h_3,…, and h_{N_p}, were separated from one another by deleting the edges between them, and a speech-unit HMM set was constructed. The model for each spoken word was represented by connecting these speech-unit HMMs.

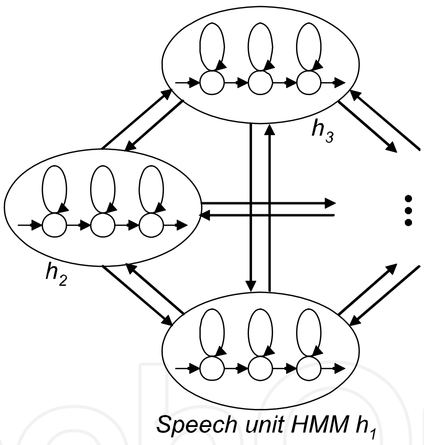


Figure 3. Structure of HMM for learning speech units

4.3 Number of speech units

In the above method, the number, N_p, of speech-unit models was determined empirically. However, ideally it should be learned from speech data. A method of learning the number of speech units and the number of words simultaneously from data comprising pairs that consist of an image of an object and a spoken word describing it has already been presented (Iwahashi, 2003a). It operates in a batch-like manner using information included in both the image and speech observations.

5. Learning Words

The lexicon consists of a set of words, and each word consists of statistical models of speech and a concept. Some words refer to objects and others refer to the motions of moving objects.

5.1 Words referring to objects

In general, the difficulty of acquiring words referring to objects can be ascribed to the difficulties involved in specifying features and applying them to other objects.

Specification: The acoustic features of a spoken word and the visual features of an object to which it refers should be specified using spatiotemporally continuous audio-visual data. For speech, this means that a continuously spoken utterance is segmented into intervals first, and then acoustic features are extracted from one of the segmented intervals. For objects, this means that an object is first selected for a given situation, and then the spatial part of the object is segmented; after that, visual features are extracted from the segmented part of the object.

Extension: To create categories for a given word and its meaning, it is necessary to determine what other features fall into the category to which the specified features belong. This extension of the features of a word's referent to form the word's meaning has been investigated through psychological experiments (Bloom, 2000). When shown an object and given a word for it, human subjects tend to extend the features of the referent immediately in order to infer a particular meaning of the word; this is a cognitive ability called fast mapping (e.g., (Imai & Gentner, 1997)), although such inference is not necessarily correct. For machines, however, the difficulty in acquiring spoken words arises not only from the difficulty in extending the features of referents but also from the difficulty in understanding spoken words. This is because machines are currently much less accurate in recognizing speech than humans; thus, it is not easy for machines to determine whether two different speech segments belong to the same word category.

The learning method described here mainly addresses the problem of extension, in which learning is carried out in an interactive way (Iwahashi, 2004). In learning, the user shows a physical object to the robot and at the same time speaks the name of the object or a word describing it. The robot then decides whether the input word is one in its vocabulary (a *known* word) or not (an *unknown* word). If the robot judges that the input word is an unknown word, it registers it in its vocabulary. If the robot judges that it cannot make an accurate decision, it asks the user a question to determine whether the input word is part of its vocabulary. For the robot to make a correct decision, it uses both speech and visual information about the objects. For example, when the user shows the robot an orange and says the word /ɔrinʒ/ even if the speech recognizer outputs an unknown word /are:ʒ/ as the first candidate, the system can modify it to the correct word /ɔrinʒ/ in the lexicon by using visual clues. Such a decision is carried out using a function that represents the confidence that an input pair of image o and speech s belongs to each existing word category w and is adaptively changed online.

Each word or lexical item to be learned includes statistical models, $p(s|w)$ and $p(o|w)$, for the spoken word and for the object image category of its meaning, respectively. The model for each image category $p(o|w)$ is represented by a Gaussian function in a twelvedimensional visual feature space (in terms of shape, colour, and size), and learned using a Bayesian method (e.g., (Degroot, 1970)) every time an object image is given.

The Bayesian method makes it possible to determine the area in the feature space that belongs to an image category in a probabilistic way, even if there is only a single sample. The model for each spoken word $p(s \mid w)$ is represented by a concatenation of speech-unit HMMs; this extends a speech sample to a spoken word category.

In experiments, forty words were successfully learned including those that refer to whole objects, shapes, colours, sizes, and combinations of these things.

5.2 Words referring to motions

While the words referring to objects are nominal, the words referring to motions are relational. The concept of the motion of a moving object can be represented by a time-varying spatial relation between a trajector and landmarks, where the trajector is an entity characterized as the figure within a relational profile, and the landmarks are entities characterized as the ground that provide points of reference for locating the trajector (Langacker, 1991). Thus, the concept of the trajectory of an object depends on the landmarks. In Fig. 4, for instance, the trajectory of the stuffed toy on the left moved by the user, as indicated by the white arrow, is understood as *move over* and *move onto* when the landmarks are considered to be the stuffed toy in the middle and the box on the right, respectively. In general, however, information about what is a landmark is not observed in learning data. The learning method must infer the landmark selected by a user in each scene. In addition, the type of coordinate system in the space should also be inferred to appropriately represent the graphical model for each concept of a motion.

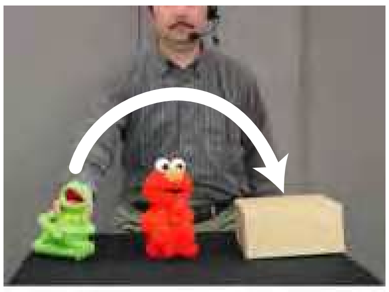


Figure 4. Scene in which utterances were made and understood

In the method for learning words referring to motions (Haoka & Iwahashi, 2000), in each episode, the user moves an object while speaking a word describing the motion. Through a sequence of such episodes, the set comprising pairs of a scene O before an action, and action a, $D_m = \{(a_1, O_1), (a_2, O_2), \dots, (a_{N_m}, O_{N_m})\}$, is given as learning data for a word referring to a motion concept. Scene O_i includes the set of positions $o^i_{j,p}$ and features $o^i_{j,f}$ concerning colour, size, and shape, $j = 1, \dots J_i$, of all objects in the scene. Action a_i is represented by a pair (t_i, u_i) consisting of trajector object t_i and trajectory u_i of its movement. The concepts regarding motions are represented by probability density functions of the trajectory u of moved objects. Four types of coordinate systems

$k \in \{1,2,3,4\}$ are considered. The probability density function $p(u \mid o_{l,p}, k_w, \lambda_w)$ for the trajectory of the motion referred to by word w is represented by HMM λ_w and the type of coordinate system k_w, given positions $o_{l,p}$ of a landmark. The HMM parameters λ_w of the motion are learned while the landmarks l and the type of coordinate system k_w are being inferred based on the EM (expectation maximization) algorithm, in which a landmark is taken as a latent variable as

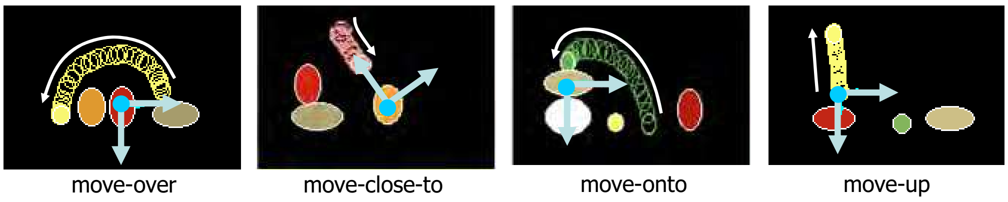
$$(\mathbf{l}, k_w, \lambda_w) = \arg\max_{\mathbf{l},k,\lambda} \sum_{i=1}^{N_m} \log p\left(u_i \,\middle|\, o^{i}_{l_i,p}, k, \lambda\right) , \tag{1}$$

where $\mathbf{l} = [l_1, l_2, \ldots, l_{N_m}]$. Here, l_i is a discrete variable across all objects in each scene O_i, and it represents a landmark object. The number of states in the HMM is also learned through cross-validation. In experiments, six motion concepts, “move-over,” “move-onto,” “move-close-to”, “move-away”, “move-up”, and “move-circle”, were successfully learned. Examples of inferred landmarks and coordinates in the learning of some motion concepts are shown in Fig. 5. A graphical model of the lexicon containing words referring to objects and motions is shown in Fig. 6.

The trajectory for the motion referred to by word w is generated by maximizing the output probability of the learned HMM, given the positions of a trajector and a landmark as

$$\tilde{u} = \arg\max_{u} p(u \mid o_{l,p}, k_w, \lambda_w) . \tag{2}$$

This maximization is carried out by solving simultaneous linear equations (Tokuda et al., 1995).



move-over

move-close-to

move-onto

move-up

Figure 5. Trajectories of objects moved in learning episodes and selected landmarks and coordinates

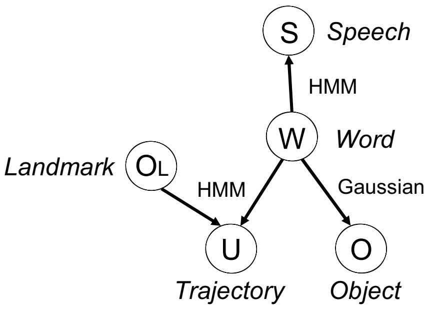


Figure 6. Graphical model of a lexicon containing words referring to objects and motions

6. Learning Grammar

To enable the robot to learn grammar, we use moving images of actions and speech describing them. The robot should detect the correspondence between a semantic structure in the moving image and a syntactic structure in the speech. However, such semantic and syntactic structures are not observable. While an enormous number of structures can be extracted from a moving image and speech, the method should select the ones with the most appropriate correspondence between them. Grammar should be statistically learned using such correspondences, and then inversely used to extract the correspondence.

The set comprising triplets of a scene O before an action, action a, and a sentence utterance s describing the action, $\mathrm{D}_g = \left\{(s_1, a_1, O_1), (s_2, a_2, O_2), ..., \left(s_{N_g}, a_{N_g}, O_{N_g}\right)\right\}$, is given in this order as learning data. It is assumed that each utterance is generated based on the stochastic grammar G based on a conceptual structure. The conceptual structure used here is a basic schema used in cognitive linguistics, and is expressed with three conceptual attributes—[motion], [trajector], and [landmark]—that are initially given to the system, and they are fixed. For instance, when the image is the one shown in Fig. 4 and the corresponding utterance is the sequence of spoken words "*large frog brown box move-onto*", the conceptual structure $z = (W_T, W_L, W_M)$ might be

$$\begin{bmatrix} \text{[trajector]} & : \textit{large frog} \\ \text{[landmark]} & : \textit{brown box} \\ \text{[motion]} & : \textit{move-onto} \end{bmatrix},$$

where the right-hand column contains the spoken word subsequences W_T, W_L, and W_M, referring to trajector, landmark, and motion, respectively, in a moving image. Let y denote the order of conceptual attributes, which also represents the order of the constituents with the conceptual attributes in an utterance. For instance, in the above utterance, the order is [trajector]-[landmark]-[motion]. The grammar is represented by the set comprising the occurrence probabilities of the possible orders as $G = \{P(y_1), P(y_2), ..., P(y_k)\}$.

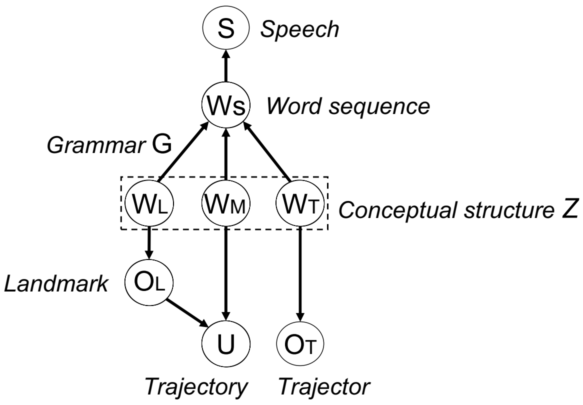


Figure 7. Graphical model of lexicon and grammar

Joint probability density function $p(s,a,O;L,G)$, where L denotes a parameter set of the lexicon, is represented by a graphical model with an internal structure that includes the parameters of grammar G and conceptual structure z that the utterance represents (Fig. 7). By assuming that $p(z,O;L,G)$ is constant, we can write the joint log-probability density function as

$$\begin{aligned}
&\log p(s,a,O;L,G) \\
&\quad = \log \sum_{z} p(s \mid z;L,G)\, p(a \mid x,O;L,G)\, p(z,O;L,G) \\
&\quad \approx \alpha \max_{z,l} \Big(\log p(s \mid z;L,G) && \text{[Speech]} \qquad (3) \\
&\qquad\quad + \log p\big(u \mid o_{l,p},W_M;L\big) && \text{[Motion]} \\
&\qquad\quad + \log p\big(o_{t,f} \mid W_T;L\big) + \log p\big(o_{l,f} \mid W_L;L\big)\Big), && \text{[Static image of object]}
\end{aligned}$$

where α is a constant value of $p(z,O;L,G)$. Furthermore, t and l are discrete variables across all objects in each moving image and represent, respectively, a trajector object and a landmark object. As an approximation, the conceptual structure $z=(W_T,W_L,W_M)$ and landmark l, which maximizes the output value of the function, are used instead of summing up for all possible conceptual structures and landmarks.

Estimate $\tilde{G}_i$ of grammar G given i th learning data is obtained as the maximum values of the posterior probability distribution as

$$\tilde{G}_i = \arg\max_{G} p\left(G \mid \mathrm{D}_g^i;L\right), \tag{4}$$

where D_g^i denotes learning sample set $\{(s_1,a_1,O_1),(s_2,a_2,O_2),\ldots,(s_i,a_i,O_i)\}$.

An utterance s asking the robot to move an object is understood using lexicon L and grammar G. Accordingly, one of the objects, t, in current scene O is grasped and moved along trajectory u by the robot. Action $\tilde{a}=(\tilde{t},\tilde{u})$ for utterance s is calculated as

$$\tilde{a} = \arg\max_{a} = \log p\left(s,a,O;L,\tilde{G}\right). \tag{5}$$

This means that from among all the possible combinations of conceptual structure z, trajector and landmark objects t and l, and trajectory u, the method selects the combination that maximizes the value of the joint log-probability density function $\log p(s,a,O;L,G)$.

7. Learning Pragmatic Capability for Situated Conversations

7.1 Difficulty

As mentioned in Sec. 1, the meanings of utterances are conveyed based on certain beliefs shared by those communicating in the situations. From the perspective of objectivity, if those communicating want to logically convince each other that proposition *p* is a shared belief, they must prove that the infinitely nested proposition, "They have information that they have information that … that they have information that *p*", also holds. However, in reality, all we can do is assume, based on a few clues, that our beliefs are identical to those of the other people we are talking to. In other words, it can never be guaranteed that our beliefs are identical to those of other people. Because shared beliefs defined from the viewpoint of objectivity do not exist, it is more practical to see shared beliefs as a process of interaction between the belief systems held by each person communicating. The processes of generating and understanding utterances rely on the system of beliefs held by each person, and this system changes autonomously and recursively through these two processes. Through utterances, people simultaneously send and receive both the meanings of their words and, implicitly, information about one another's systems of beliefs. This dynamic process works in a way that makes the belief systems consistent with each other. In this sense, we can say that the belief system of one person couples structurally with the belief systems of those with whom he or she is communicating (Maturana, 1978).

When a participant interprets an utterance based on their assumptions that certain beliefs are shared and is convinced, based on certain clues, that the interpretation is correct, he or she gains the confidence that the beliefs are shared. On the other hand, since the sets of beliefs assumed to be shared by participants actually often contain discrepancies, the more beliefs a listener needs to understand an utterance, the greater the risk that the listener will misunderstand it.

As mentioned above, a pragmatic capability relies on the capability to infer the state of a user's belief system. Therefore, the method should enable the robot to adapt its assumption of shared beliefs rapidly and robustly through verbal and nonverbal interaction. The method should also control the balance between (i) the transmission of the meaning of utterances and (ii) the transmission of information about the state of belief systems in the process of generating utterances.

The following is an example of generating and understanding utterances based on the assumption of shared beliefs. Suppose that in the scene shown in Fig. 4 the frog on the left has just been put on the table. If the user in the figure wants to ask the robot to move a frog onto the box, he may say, "*frog box move-onto*". In this situation, if the user assumes that the robot shares the belief that the object moved in the previous action is likely to be the next target for movement and the belief that the box is likely to be something for the object to be moved onto, he might just say "*move-onto*"[1]. To understand this fragmentary and ambiguous utterance, the robot must possess similar beliefs. If the user knows that the robot has responded by doing what he asked it to, this would strengthen his confidence that the beliefs he has assumed to be shared really are shared. Conversely, when the robot wants to ask the user to do something, the beliefs that it assumes to be shared are used in the same way. It can be seen that the former utterance is more effective than the latter in transmitting the meaning of the utterance, while the latter is more effective in transmitting information about the state of belief systems.

[1] Although the use of a pronoun might be more natural than the deletion of noun phrases in some languages, the same ambiguity in meaning exists in both such expressions.

7.2 Representation of a belief system

To cope with the above difficulty, the belief system of the robot needs to have a structure that reflects the state of the user's belief system so that the user and the robot infer the state of each other's belief systems. This structure consists of the following two parts:

The shared belief function represents the assumption of shared beliefs and is composed of a set of belief modules with values (local confidence values) representing the degree of confidence that each belief is shared by the robot and the user.

The global confidence function represents the degree of confidence that the whole of the shared belief function is consistent with the shared beliefs assumed by the user.

Such a belief system is depicted in Fig. 8. The beliefs we used are those concerning speech, motions, static images of objects, and motion-object relationship, and the effect of behavioural context. The motion-object relationship and the effect of behavioural context are represented as follows.

Motion-object relationship $B_R\left(o_{t,f}, o_{l,f}, W_M; R\right)$: The motion-object relationship represents the belief that in the motion corresponding to motion word W_M , feature $o_{t,f}$ of object t and feature $o_{l,f}$ of object l are typical for a trajector and a landmark, respectively. This belief is represented by a conditional multivariate Gaussian probability density function, $p\left(o_{t,f}, o_{l,f} \mid W_M; R\right)$, where R is its parameter set.

Effect of behavioural context $B_H\left(i, q; H\right)$: The effect of behavioural context represents the belief that the current utterance refers to object i , given behavioural context q . Here, q includes information on which objects were a trajector and a landmark in the previous action and which object the user's current gesture refers to. This belief is represented by a parameter set H .

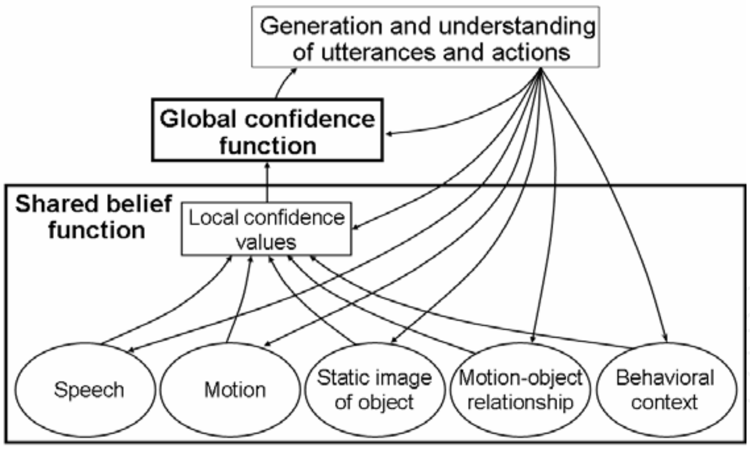

Generation and understanding of utterances and actions
Global confidence function
Shared belief function
Local confidence values
Speech
Motion
Static image of object
Motion-object relationship
Behavioral context

Figure 8. Belief system of robot that consists of shared belief and global confidence functions

7.3 Shared belief function

The beliefs described above are organized and assigned local confidence values to obtain the shared belief function used in the processes of generating and understanding utterances.

This shared belief function Ψ is the extension of $\log p(s,a,O;L,G)$ in Eq. 3. The function outputs the degree of correspondence between utterance s and action a. It is written as

$$\begin{aligned}
&\Psi(s,a,O,q,L,G,R,H,\Gamma) \\
&\quad = \max_{z,l}\Big(\gamma_1 \log p(s \mid z;L,G) && \text{[Speech]} \\
&\qquad + \gamma_2 \log p\left(u \mid o_{l,p}, W_M; L\right) && \text{[Motion]} \\
&\qquad + \gamma_2 \left(\log p\left(o_{t,f} \mid W_T; L\right) + \log p\left(o_{l,f} \mid W_L; L\right)\right) && \text{[Static image of object]} \quad (6) \\
&\qquad + \gamma_3 \log p\left(o_{t,f}, o_{l,f} \mid W_M; R\right) && \text{[Motion-object relationship]} \\
&\qquad + \gamma_4 \left(B_H(t,q;H) + B_H(l,q;H)\right)\Big), && \text{[Behavioural context]}
\end{aligned}$$

where $\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ is a set of local confidence values for beliefs corresponding to the speech, motion, static images of objects, motion-object relationship, and behavioural context. Given O, q, L, G, R, H, and Γ, the corresponding action $\tilde{a} = (\tilde{t}, \tilde{u})$, understood to be the meaning of utterance s, is determined by maximizing the shared belief function as

$$\tilde{a} = \arg\max_{a} \Psi(s,a,O,q,L,G,R,H,\Gamma). \tag{7}$$

7.4 Global confidence function

The global confidence function f outputs an estimate of the probability that the robot's utterance s will be correctly understood by the user . It is written as

$$f(d) = \frac{1}{\pi}\arctan\left(\frac{d-\lambda_1}{\lambda_2}\right) + 0.5, \tag{8}$$

where λ_1 and λ_2 are the parameters of the function and input d of this function is a margin in the value of the output of the shared belief function between an action that the robot asks the user to take and other actions in the process of generating an utterance. Margin d in generating utterance s to refer to action a in scene O in behavioural context q is defined as

$$d(s,a,O,q,L,G,R,H,\Gamma) = \Psi(s,a,O,q,L,G,R,H,\Gamma) - \max_{A \neq a} \Psi(s,A,O,q,L,G,R,H,\Gamma). \tag{9}$$

Examples of the shapes of global confidence functions are shown in Fig. 9. Clearly, a large margin increases the probability of the robot being understood correctly by the user. If there is a high probability of the robot's utterances being understood correctly even when the margin is small, it can be said that the robot's beliefs are consistent with those of the user. The example of a shape of such a global confidence function is indicated by "strong". In contrast, the example of a shape when a large margin is necessary to get a high probability is indicated by "weak".

When the robot asks for action a in scene O in behavioural context q, it generates utterance $\tilde{s}$ so as to bring the value of the output of f as close as possible to the value of parameter ξ, which represents the target probability of the robot's utterance being understood correctly. This utterance can be represented as

$$\tilde{s} = \arg\min_{s} \left| f\left(d\left(s,a,O,q,L,G,R,H,\Gamma\right)\right) - \xi \right| . \tag{10}$$

The robot can increase its chance of being understood correctly by using more words. On the other hand, if the robot can predict correct understanding with a sufficiently high probability, it can manage with a fragmentary utterance using a small number of words.

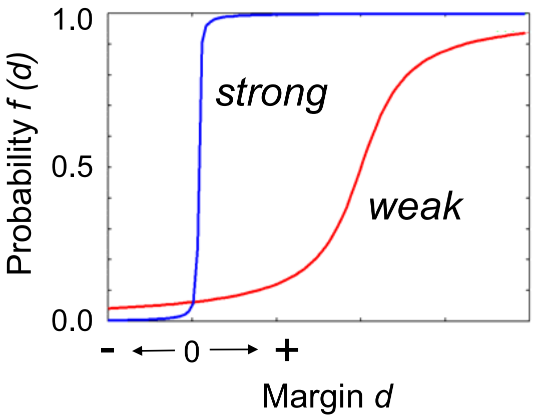


Figure 9. Examples of shapes of global confidence functions

7.5 Learning methods

The shared belief function Ψ and the global confidence function f are learned separately in the processes of utterance understanding and utterance generation by the robot, respectively.

7.5.1 Utterance understanding by the robot

Shared belief function Ψ is learned incrementally, online, through a sequence of episodes, each of which comprises the following steps.

1. Through an utterance and a gesture, the user asks the robot to move an object.
2. The robot acts on its understanding of the utterance.
3. If the robot acts correctly, the process ends. Otherwise, the user slaps its hand.
4. The robot acts in a different way.
5. If the robot acts incorrectly, the user slaps its hand. The process ends.

In each episode, a quadruplet (s,a,O,q) comprising the user's utterance s, scene O, behavioural context q, and action a that the user wants to ask the robot to take, is used. The robot adapts the values of parameter set R for the belief about the motion-object relationship, parameter set H for the belief about the effect of the behavioural context, and local confidence parameter set Γ. Lexicon L and grammar G were learned beforehand, as described in the previous sections. When the robot acts correctly in the first or second trials, it learns R by applying the Bayesian learning method using the information about features of trajector and landmark objects $o_{t,f}$, $o_{l,f}$ and motion word W_M in the utterances. In addition, when the robot acts correctly in the second trial, it associates utterance s, correct action a, incorrect action A from the first trial, scene O, and behavioural context q with one another and makes these associations into a learning sample. When the ith sample $(s_i, a_i, A_i, O_i, q_i)$ is obtained based on this process of association, H_i and Γ_i are adapted to approximately minimize the probability of misunderstanding as

$$\left(\tilde{H}_i, \tilde{\Gamma}_i\right) = \arg\min_{H,\Gamma} \sum_{j=i-K}^{i} w_{i-j} g\left(\Psi\left(s_j, a_j, O_j, q_j, L, G, R_i, H, \Gamma\right) - \Psi\left(s_j, A_j, O_j, q_j, L, G, R_i, H, \Gamma\right)\right), \quad (11)$$

where $g(x)$ is $-x$ if $x < 0$ and 0 otherwise, and K and w_{i-j} represent the number of latest samples used in the learning process and the weights for each sample, respectively.

7.5.2 Utterance generation by the robot

Global confidence function f is learned incrementally, online through a sequence of episodes, each of which consists of the following steps.

1. The robot generates an utterance to ask the user to move an object.
2. The user acts according to his or her understanding of the robot's utterance.
3. The robot determines whether the user's action is correct.

In each episode, a triplet (a, O, q) comprising scene O, behavioural context q, and action a that the robot needs to ask the user to take is provided to the robot before the interaction. The robot generates an utterance that brings the value of the output of global confidence function f as close to ξ as possible. After each episode, the value of margin d in the utterance generation process is associated with information about whether the utterance was understood correctly, and this sample of associations is used for learning. The learning is done online and incrementally so as to approximate the probability that an utterance will be understood correctly by minimizing the weighted sum of squared errors in the most recent episodes. After the ith episode, parameters λ_1 and λ_2 are adapted as

$$\left[\lambda_{1,i}, \lambda_{2,i}\right] \leftarrow (1-\delta)\left[\lambda_{1,i-1}, \lambda_{2,i-1}\right] + \delta\left[\tilde{\lambda}_{1,i-1}, \tilde{\lambda}_{2,i-1}\right], \quad (12)$$

where

$$\left(\tilde{\lambda}_{1,i}, \tilde{\lambda}_{2,i}\right) = \arg\min_{\lambda_1,\lambda_2} \sum_{j=i-K}^{i} w_{i-j} \left(f\left(d_j; \lambda_1, \lambda_2\right) - e_j\right)^2, \quad (13)$$

where e_i is 1 if the user's understanding is correct and 0 if it is not, and δ is the value that determines learning speed.

7.6 Experimental results

7.6.1 Utterance understanding by the robot

At the beginning of the sequence for learning shared belief function Ψ , the sentences were relatively complete (e.g., "*green frog red box move-onto*"). Then the lengths of the sentences were gradually reduced (e.g., "*move-onto*") to become fragmentary so that the meanings of the sentences were ambiguous. At the beginning of the learning process, the local confidence values γ_1 and γ_2 for speech, static images of objects, and motions were set to 0.5 , while γ_3 and γ_4 were set to 0 .

R could be estimated with high accuracy during the episodes in which relatively complete utterances were given and understood correctly. In addition, H and Γ could be effectively estimated based on the estimation of R during the episodes in which fragmentary utterances were given. Figure 10 shows changes in the values of γ_1 , γ_2 , γ_3 , and γ_4 . The values did not change during the first thirty-two episodes because the sentences were relatively complete and the actions in the first trials were all correct. Then, we can see that value γ_1 for speech decreased adaptively according to the ambiguity of a given sentence, whereas the values γ_2 , γ_3 , and γ_4 for static images of objects, motions, the motion-object relationship, and behavioural context increased. This means that non-linguistic information was gradually being used more than linguistic information.

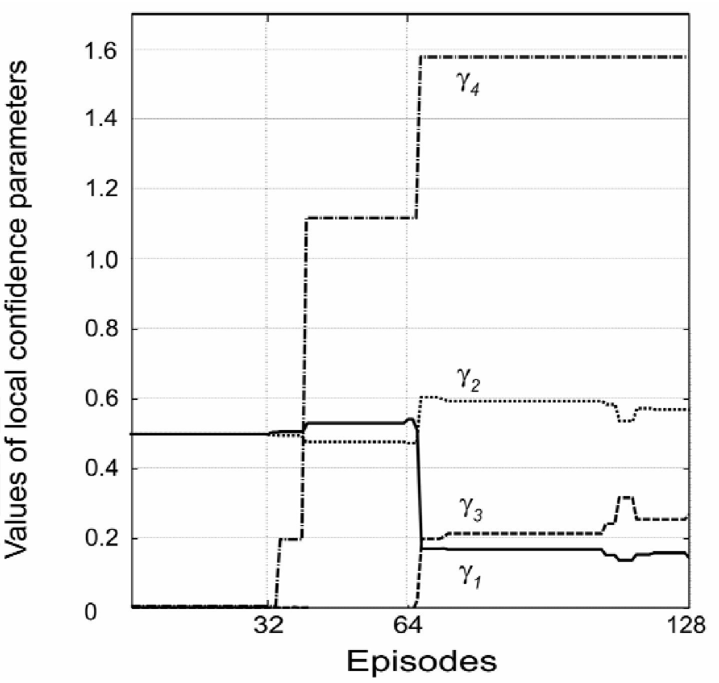
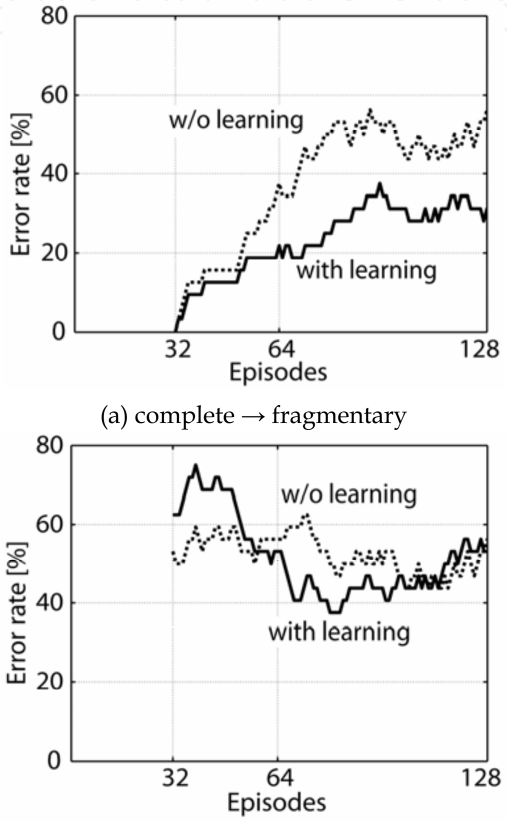


Figure 10. Changes in values of local confidence parameters

Figure 11(a) shows the decision error (misunderstanding) rates obtained during the course of the interaction, along with the error rates obtained for the same learning data by keeping the values of the parameters of the shared belief function fixed to their initial values. We can see that the learning was effective. In contrast, when fragmentary utterances were provided over the whole sequence of the interaction, the robot did not learn well (Fig. 11(b)) because it misunderstood the utterances too often.

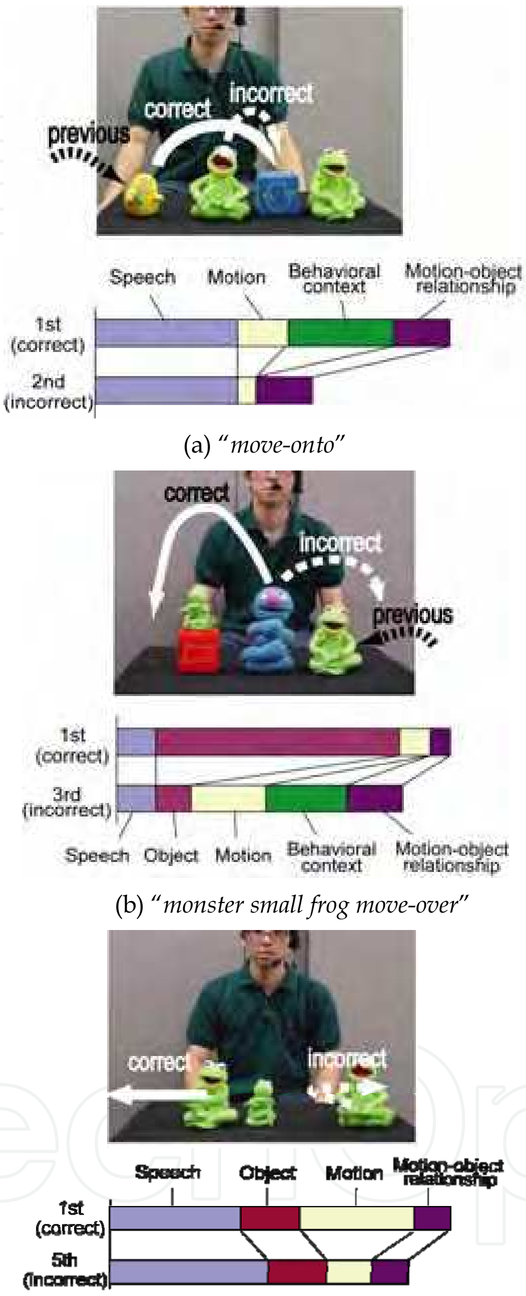


(a) complete → fragmentary

(b) fragmentary → fragmentary

Figure 11. Change in decision error rate

Examples of actions generated as a result of correct understanding are shown together with the output log-probabilities from the weighted beliefs in Figs. 12(a), (b), and (c) along with the second, third, and fifth choices for action, respectively, which were incorrect. It is clear that each non-linguistic belief was used appropriately in understanding the utterances according to their relevance to the situations. The beliefs about the effect of the behavioural context were more effective in Fig. 12(a) , while in Fig. 12(b), the beliefs about the concepts of the static images of objects were more effective than other non-linguistic beliefs in leading to the correct understanding. In Fig. 12(c), even when error occurred in speech recognition, the belief about the motion concept was effectively used to understand the utterance with ambiguous meaning.



(a) "*move-onto*"

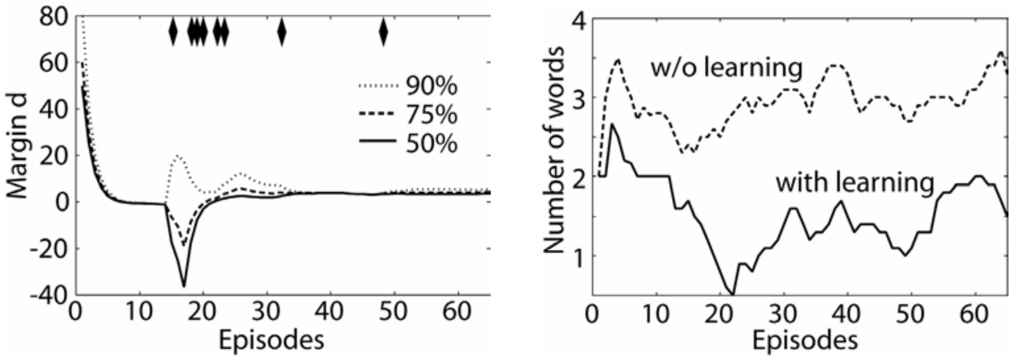
(b) "*monster small frog move-over*"

(c) *"large frog small frog move-away"*

Figure 12. Examples of actions generated as a result of correct understanding and weighted output log-probabilities from beliefs, along with second, third, and fifth action choices

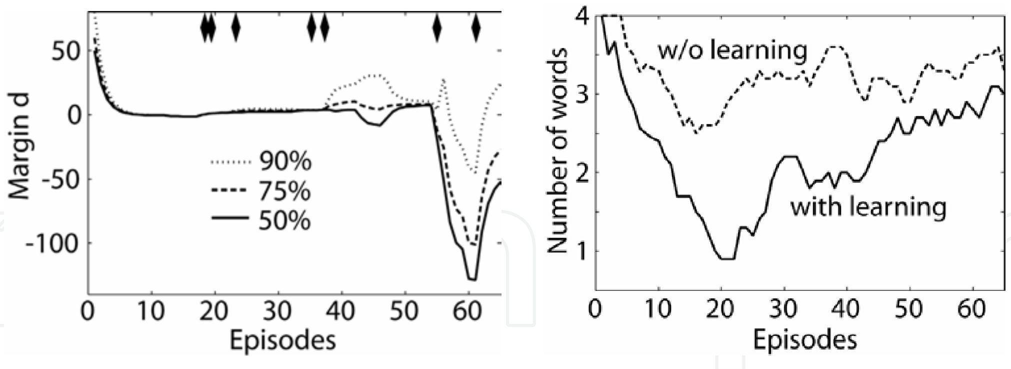
7.6.2 Utterance generation by the robot

In each episode for learning global confidence function f, the robot generated an utterance so as to make the value of the output of the global confidence function as close to $\xi = 0.75$ as possible. Even when the target value ξ was fixed at 0.75, we found that the obtained values were widely distributed around it. The initial shape of the global confidence function was set to make $f^{-1}(0.9) = 161$, $f^{-1}(0.75) = 120$, and $f^{-1}(0.5) = 100$, meaning that a large margin was necessary for an utterance to be understood correctly. In other words, the shape of f in this case represents weak confidence. Note that when all of the values are close to 0, the slope in the middle of f is steep, and the robot makes the decision that a small margin is sufficient for its utterances to be understood correctly. The shape of f in this case represents strong confidence.



(a) (b)

Figure 13. (a) Changes in global confidence function and (b) number of words needed to describe objects in each utterance, $\xi = 0.75$



(a) (b)

Figure 14. (a) Changes in global confidence function and (b) number of words needed in each utterance to describe objects, $\xi = 0.95$

The changes in $f(d)$ are shown in Fig. 13(a), where three lines have been drawn for $f^{-1}(0.9)$, $f^{-1}(0.75)$, and $f^{-1}(0.5)$ to make the shape of f easily recognizable. The interactions in which utterances were misunderstood are depicted in the upper part of the graph by the black diamonds. Figure 13(b) displays the changes in the moving average of the number of words used to describe the objects in each utterance, along with the changes obtained when f was not learned, which are shown for comparison. After the learning began, the slope in the middle of f rapidly became steep, and the number of words uttered decreased. The function became temporarily unstable with $f^{-1}(0.5) < 0$ at around the 15 th episode. The number of words uttered then became too small, which sometimes led to misunderstanding. We might say that the robot was overconfident in this period. Finally, the slope became steep again at around the 35th episode.

We conducted another experiment in which the value of parameter ξ was set to 0.95. The result of this experiment are shown in Fig. 14. It is clear that, after approximately the 40th interaction, the change in f became very unstable and the number of words became large. We found that f became highly unstable when the utterances with a large margin d were not understood correctly.

8. Discussion

8.1 Sharing the risk of being misunderstood

The experiments in learning a pragmatic capability illustrate the importance of sharing the risk of not being understood correctly between the user and the robot.

In the learning period for utterance understanding by the robot, the values of the local confidence parameters changed significantly when the robot acted incorrectly in the first trial and correctly in the second trial. To facilitate learning, the user had to gradually increase the ambiguity of utterances according to the robot's developing ability to understand them and had to take the risk of not being understood correctly. In the its learning period for utterance generation, the robot adjusted its utterances to the user while learning the global confidence function. When the target understanding rate ξ was set to 0.95, the global confidence function became very unstable in cases where the robot's expectations of being understood correctly at a high probability were not met. This instability could be prevented by using a lower value of ξ, which means that the robot would have to take a greater risk of not being understood correctly.

Accordingly, in human-machine interaction, both users and robots must face the risk of not being understood correctly and thus adjust their actions to accommodate such risk in order to effectively couple their belief systems. Although the importance of controlling the risk of error in learning has generally been seen as an exploration-exploitation trade-off in the field of reinforcement learning by machines (e.g., (Dayan & Sejnowski, 1996)), we argue here that the mutual accommodation of the risk of error by those communicating is an important basis for the formation of mutual understanding.

8.2 Incomplete observed information and fast adaptation

In general, an utterance does not contain complete information about what a speaker wants to convey to a listener. The proposed learning method interpreted such utterances according to the situation by providing necessary but missing information by making use of the assumption of shared beliefs. The method also enabled the robot and the user to adapt such an assumption of shared beliefs to each other with little interaction. We can say that the method successfully coped with two problems faced by systems interacting with the physical world: the incompleteness of observed information and fast adaptation (Matsubara & Hashida, 1989).

Some previous studies have been done in terms of these problems. In the field of autonomous robotics, the validity of the architecture in which sub-systems are allocated in parallel has been demonstrated (Brooks, 1986). This, however, failed to be applied to large-scale systems because of the lack of a mathematical theory for the interaction among the sub-systems. On the other hand, Bayesian networks (Pearl, 1988) have been studied intensively, providing a probabilistic theory of the interaction among the sub-systems. This can cope with the incompleteness of observed information in large-scale systems but does not address fast adaptation.

Shared belief function Ψ, which is a large-scale system, has the merits of both these approaches. It is a kind of Bayesian network with statistical models of beliefs allocated in parallel in which weighting values Γ are added to these models, as shown in Eq. 6. Based on both the parallel allocation of sub-systems and the probabilistic theory, this method can cope successfully with the incompleteness of observed information and can achieve fast adaptation of the function by changing weighting values.

8.3 Initial setting

The No Free Lunch Theory (Wolpert, 1995) shows that when no prior knowledge about a problem exists, it is not possible to assume that one learning algorithm is superior to another. That is, there is no learning method that is efficient for all possible tasks. This suggests that attention should be paid to domain specificity as well as versatility.

In the methods described here, the initial setting for the learning was chosen by taking into account the generality and efficiency of language learning. The conceptual attributes—[motion], [trajector], and [landmark]—were given beforehand because they are general and essential in linguistic and other cognitive processes. Given this setting, however, the constructions that the method could learn were limited to those like transitive and ditransitive ones. In future work, we will study how to overcome this limitation.

8.4 Abstract meanings

The image concepts of objects that are learned by the methods described in Sec. 5 are formed directly from perceptual information. However, we must consider words that refer to concepts that are more abstract and that are not formed directly from perceptual information, such as "tool," "food," and "pet". In a study on the abstract nature of the meanings of symbols (Savage-Rumbaugh, 1986), it was found that chimpanzees could learn the lexigrams (graphically represented words) that refer to both individual object categories (e.g., "banana", "apple", "hammer", and "key") and the functions ("tool" and "food") of the objects. They could also learn the connection between the lexigrams referring to these two kinds of concepts and generalize it appropriately to connect new lexigrams for individual objects to lexigrams for functions.

A method enabling robots to gain this ability has been proposed (Iwahashi et al., 2006). In that method, the motions of objects are taken as their functions. The main problem is the decision regarding whether the meaning of a new input word applies to a concept formed directly from perceptual information or to a function of objects. Because these two kinds of concepts are allocated to the states of different nodes in the graphical model, the problem becomes the selection of the structures of the graphical model. This selection is made by the Bayesian principle with the calculation of posterior probabilities using the variational Bayes method (Attias, 1999).

8.5 Prerequisites for conversation

Language learning can be regarded as a kind of role reversal imitation (Carpenter et al., 2005). To coordinate roles in a joint action among participants, the participants should be able to read the intentions of the others. It is known that at a very early stage of development, infants become able to understand the intentional actions of others (Behne et al., 2005) and even to understand that others might have beliefs different from their own (Onishi & Baillargeon, 2005).

In the method described in this chapter, the robot took the speech acts of input utterances as descriptive or directive. If the utterance was descriptive in terms of the current situation, the robot learned a lexicon or grammar. If the utterance was directive, the robot moved an object. The distinction between descriptive and directive acts was made by taking account of both the user's behaviour and speech. The simple rule for this distinction was given to the robot and the user beforehand, so they knew it.

A learning method that enables robots to understand the kinds of speech act in users' utterances has been presented (Taguchi et al., 2007). Using this method, robots came to understand their roles in interactions by themselves based on role reversal imitation and came to learn such distinction of speech acts which requests for actions. Eventually the robot came to respond to a request by moving an object, and to answer a question by speaking and pointing. The method was developed by expanding the graphical model described here.

8.6 Psychological investigation

The experimental results showed that the robot could learn new concepts and form a system of beliefs that it assumed the user also had. Because the user and the robot came to understand fragmentary and ambiguous utterances, they must have shared similar beliefs and must have been aware that they shared them. It would be interesting to investigate through psychological experiments the dynamics of belief sharing between users and robots.

9. Conclusion

A developmental approach to language processing for situated human-robot conversations was presented. It meets three major requirements that existing language processing methods cannot: grounding, scalability, and sharing of beliefs. The proposed method enabled a robot to learn language communication capability online with relatively little verbal and nonverbal interaction with a user by combining speech, visual, and behavioural reinforcement information in a probabilistic framework. The adaptive behaviour of the belief systems is modelled by the structural coupling of the belief systems held by the robot and the user, and it is executed through incremental online optimisation during the process of interaction. Experimental results revealed that through a small, but practical number of learning episodes with a user, the robot was eventually able to understand even fragmentary and ambiguous utterances, act upon them, and generate utterances appropriate for the given situation. Future work includes investigating the learning of more complex linguistic knowledge and learning in a more natural environment.

10. Acknowledgements

This work was supported by a research grant from the National Institute of Informatics.

11. References

Allen, J.; Byron, D.; Dzikovska, M.; Ferguson, G.; Galescu, L. & Stent, A. (2001). Toward conversational human-computer interaction. *AI Magazine*, Vol. 22, Issue. 4, pp. 27-38

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of Int. Conf. on Uncertainty in Artificial Intelligence*, pp. 21–30

Baum, L. E.; Petrie, T.; Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164–171

Behne, T.; Carpenter, M.; Call, J. & Tomasello, M. (2005). Unwilling versus unable – infants' understanding of intentional action. *Developmental Psychology*, Vol. 41, No. 2, pp. 328–337

Bloom, P. (2000). How children learn the meanings of words. MIT Press

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, Vol. 61, pp. 1–61

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, Vol. 1, pp.14–23

Carpenter, M.; Tomasello, M.; Striano, T. (2005). Role reversal imitation and language in typically developing infants and children with autism. *INFANCY*, Vol. 8, No. 3, pp. 253–278

Clark, H. (1996). Using Language. Cambridge University Press

Dayan, P. & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, Vol. 25, pp. 5–22

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, pp. 357–366

DeGroot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill

Dyer, M. G. & Nenov, V. I. (1993). Learning language via perceptual/motor experiences. *Proceedings of Annual Conf. of the Cognitive Science Society*, pp. 400–405

Gorin, A.; Levinson, S. & Sanker, A. (1994). An experiment in spoken language acquisition. *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 1, pp. 224–240

Haoka, T. & Iwahashi, N. (2000). Learning of the reference-point-dependent concepts on movement for language acquisition. *Tech. Rep. of the Institute of Electronics, Information and Communication Engineers PRMU2000-105*, pp.39-45

Imai, M. & Gentner, D. (1997). A crosslinguistic study of early word meaning – universal ontology and linguistic influence. *Cognition*, Vol. 62**,** pp. 169–200

Iwahashi, N. (2003a). Language acquisition through a human-robot interface by combining speech, visual, and behavioral information. *Information Sciences*, Vol. 156, pp. 109-121

Iwahashi, N. (2003b). A method of coupling of belief systems through human-robot language interaction. *Proceedings of IEEE Workshop on Robot and Human Interactive Communication*, pp. 385-390

Iwahashi, N. (2004). Active and unsupervised learning of spoken words through a multimodal interface. *Proceedings of IEEE Workshop on Robot and Human Interactive Communication*, pp. 437-442

Iwahashi, N.; Satoh, K. & Asoh, H. (2006). Learning abstract concepts and words from perception based on Bayesian model selection. *Tech. Rep. of the Institute of Electronics, Information and Communication Engineers PRMU-2005-234*, pp. 7-12

Jordan, M. I. & Sejnowski, T.J. Eds. (2001). Graphical Models - Foundations of Neural Computation. The MIT Press

Langacker, R. (1991). Foundation of cognitive grammar. Stanford University Press, CA

Matsubara, H. & Hashida, K. (1989). Partiality of information and unsolvability of the frame problem. *Japanese Society for Artificial Intelligence*, Vol. 4, No. 6, pp. 695–703

Maturana, H. R. (1978). Biology of language – the epistemology of reality. In: *Psychology and Biology of Language and Thought – Essay in Honor of Eric Lenneberg*, Miller, G.A., Lenneberg, E., (Eds.), pp.27–64, Academic Press

Nakagawa, S. & Masukata, M. (1995). An acquisition system of concept and grammar based on combining with visual and auditory information. *Trans. Information Society of Japan*, Vol. 10, No. 4, pp. 129–137

Onishi, K. H. & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, Vol. 308, pp. 225–258

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of Plausible Inference. Morgan Kaufmann

Persoon, E., Fu, K. S. (1977). Shape discrimination using Fourier descriptors. *IEEE Trans Systems, Man, and Cybernetics*, Vol, 7, No. 3, pp. 170–179

Regier, T. (1997). The Human Semantic Potential. MIT Press

Roy, D. (2005). Semiotic Schemas: A Framework for Grounding Language in Action and Perception. *Artificial Intelligence*, Vol. 167, Issues. 1-2, pp. 170-205

Roy, D. & Pentland, A. (2002). Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science*, Vol. 26, No. 1, pp. 113-146

Shapiro, C. S.; Ismail, O.; Santore, J. F. (2000). Our dinner with Cassie. Proceedings of AAAI 2000 Spring Symposium on Natural Dialogues with Practical Robotic Devices, pp.57–61

Steels, L. (2003). Evolving grounded communication for robots. *Trends in Cognitive Science*, Vol. 7, No. 7, pp. 308–312

Steels, L. & Kaplan, K. (2001). Aibo's first words: the social learning of language and meaning. *Evolution of Communication*, Vol. 4, No. 1, pp.3–32

Savage-Rumbaugh, E. S. (1986). Ape Language – From Conditional Response to Symbol. Columbia Univ. Press

Sperber, D. & Wilson, D. (1995). Relevance (2nd Edition). Blackwell

Taguchi, R.; Iwahashi, N. & Nitta, T. (2007). Learning of question and answer reflecting scenes of the real world. *Tech. Rep. of Japanese Society of Artificial Intelligence*, SIG-SLUD-A603-04, pp. 15-20

Tokuda, K.; Kobayashi, T. & Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In: Proceedings Int. Conf. on Acoustics, Speech and Signal Processing, pp. 660–663

Traum, D. R. (1994). A computational theory of grounding in natural language conversation. Doctoral dissertation, University of Rochester

Winograd, T. (1972). Understanding Natural Language. Academic Press New York

Wolpert, D. H. (1995). The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In Wolpert, D.H., ed.: The mathematics of Generalization, pp. 117-214, Addison-Wesley, Reading, MA

Human Robot Interaction

Edited by Nilanjan Sarkar

ISBN 978-3-902613-13-4
Hard cover, 522 pages
Publisher I-Tech Education and Publishing
Published online 01, September, 2007
Published in print edition September, 2007

Human-robot interaction research is diverse and covers a wide range of topics. All aspects of human factors and robotics are within the purview of HRI research so far as they provide insight into how to improve our understanding in developing effective tools, protocols, and systems to enhance HRI. For example, a significant research effort is being devoted to designing human-robot interface that makes it easier for the people to interact with robots. HRI is an extremely active research field where new and important work is being published at a fast pace. It is neither possible nor is it our intention to cover every important work in this important research field in one volume. However, we believe that HRI as a research field has matured enough to merit a compilation of the outstanding work in the field in the form of a book. This book, which presents outstanding work from the leading HRI researchers covering a wide spectrum of topics, is an effort to capture and present some of the important contributions in HRI in one volume. We hope that this book will benefit both experts and novice and provide a thorough understanding of the exciting field of HRI.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821