Active Learning for Generating Motion and Utterances in Object Manipulation Dialogue Tasks *

Komei Sugiura, Naoto Iwahashi, Hisashi Kawai and **Satoshi Nakamura**
National Institute of Information and Communications Technology,
3-5 Hikaridai, Seika, Soraku, Kyoto 619-0289, Japan

Abstract

In an object manipulation dialogue, a robot may misunderstand an ambiguous command from a user, such as "Place the cup down (on the table), " potentially resulting in an accident. Although making confirmation questions before all motion execution will decrease the risk of this failure, the user will find it more convenient if confirmation questions are not made under trivial situations. This paper proposes a method for estimating ambiguity in commands by introducing an active learning framework with Bayesian logistic regression to human-robot spoken dialogue. We conducted physical experiments in which a user and a manipulator-based robot communicated using spoken language to manipulate objects.

1 Introduction

For practical reasons, most dialogue management mechanisms adopted for service robots process verbal (user's utterances) and nonverbal (e.g., vision, motion and context) information separately. With these mechanisms, neither the situation nor previous experiences are taken into account when a robot processes an utterance, so there is a possibility that it will execute motions that the user had not imagined. In this study, we define "motion failure" as occurring when a robot has executed an undesirable motion because of a recognition error.

The goal of this study is to decrease the risk of motion failure. A simple solution to decrease the risk of motion failure is to make confirmation utterances before motion execution, such as "You said 'Bring me the cup.' Is this correct?" However, there are two main hurdles to generating confirmation utterances: *whether to confirm* and *how to confirm*.

The problem of *whether to confirm* is a decision-making problem of whether a confirmation utterance should be made or not. Although making confirmation utterances before all motion executions would be simple and effective, this would however seriously disrupt the dialogue. Specifically, the user will find it more convenient if confirmation questions are not made under trivial situations. In the field of spoken dialogue systems, the *whether-to-confirm* problem receives considerable attention in the context of error handling (Komatani and Kawahara 2000, Bohus and Rudnicky 2005).

The problem of *how to confirm* is the problem of paraphrasing user's commands. The sentence "Bring me a cup" is ambiguous when there are multiple cups, and asking a confirmation question such as "Do you mean the blue cup?" can disambiguate the sentence. Moreover, when direct and/or indirect objects are omitted (object ellipsis) in the user's utterance, such as "Place the cup down (on the table)," it would be preferable to generate an appropriate description of the objects. The *how-to-confirm* problem deals with the mapping between language and physical/virtual objects, and has been widely explored in Natural Language Generation (NLG) studies (e.g., Dale and Reiter 1995, Jordan and Walker 2005, Funakoshi et al.2009) presents a model for priming speech recognition using visual and contextual information.

The robotics community has recently been paying greater attention to the mapping between language and real-world information, mainly focusing on motion (Kruger et al. 2007, Sugita and Tani 2005, Inamura et al. 2004). (Ogata et al. 2007) presents an application of recurrent neural networks to the problem of handling many-to-many relationships between motion sequences and linguistic sequences. In (Takano and Nakamura 2009), a linguistic model based on the symbolization of motion patterns is proposed. Moreover, we have proposed a robot language acquisition framework "LCore" that integrates multimodal information such as speech, motion, and visual information(Iwahashi 2007).

In this study, we extend LCore with a scheme of dialogue management method based upon an adaptive confidence measure. called the *integrated confidence measure* (ICM) function. The proposed method has three key features:

1. A user model corresponding to each modality is assumed to be shared by the user and robot. This assumption enables us to introduce an active learning framework into human-robot dialogue. The user model is explained in Section 3.

2. Active learning is used for selecting the optimal utterances to generate, which effectively train the ICM function. The introduction of active learning is evaluated us-

*This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 20500186, 2008, and the National Institute of Informatics.

ing likelihood criteria in Section 5.

3. Bayesian logistic regression (BLR)(Genkin, Lewis, and Madigan 2007) is used for learning the ICM function that enables us to estimate the probability that the user's utterances will be successfully understood from multimodal information.

2 Task Environment

Object Manipulation Dialogue Task

Figure 3 shows the task environment used in this study. A user sits in front of a robot and commands the robot by speech to manipulate objects on the table located between the robot and the user. The robot is also able to command the user by speech to manipulate the objects. The objects used in the experiments are shown in Figure 2.

We assume that linguistic knowledge (e.g., phoneme sequence and, word sequence) and non-linguistic knowledge (e.g., motion and, visual information) are learned by using LCore(Iwahashi 2007). This knowledge is not given by the designer, but is learned through interaction with users. Knowledge representation in LCore is explained in Section 3 in detail. The main functions given by the designer are object extraction and calculation of visual features.

The task has three phases:

1. Robot command phase (learning phase (a))
   The robot commands the user to manipulate objects. The ICM function is trained using the proposed method.
2. User command phase (learning phase (b))
   The user commands the robot to manipulate objects. The ICM function is trained with initialization based on the results of the robot command phase.
3. Motion execution phase
   The user commands the robot to manipulate objects, however the ICM function is not updated. Motion and confirmation utterances are generated by the method proposed in (Sugiura et al. 2009).

Figure 1 shows an example of the user command phase. The figure depicts a camera image in which the robot is told to place Object 1 (Barbabright) on Object 2 (red box). The solid line shows the trajectory intended by the user. The relative trajectory between the trajector (moved object) and the reference object is modeled with a hidden Markov model (HMM)(Sugiura and Iwahashi 2007). The reference object can be the trajectory itself or a landmark characterizing the trajectory of the trajector. In the case shown in Figure 1, the trajector, reference object, and reference point are Object 1, Object 2, and Object 2's center of gravity, respectively.

Robotic Platform

Figure 3 shows the robot used in this study. The robot consists of a manipulator with seven degrees of freedom (DOFs), a four-DOF multifingered grasper, a microphone/speaker, a stereo vision camera, 3D time-of-flight camera (SR-4000), and a gaze-expression unit. Teaching signals can be provided by hitting a touch sensor on the grasper.

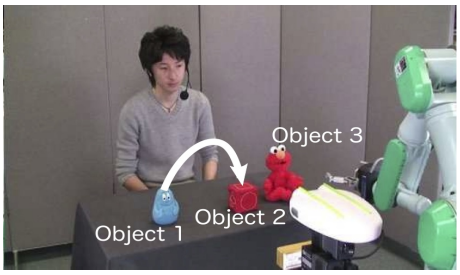


Figure 1: An example of object manipulation dialogue tasks.

Figure 2: Objects used in experiments.

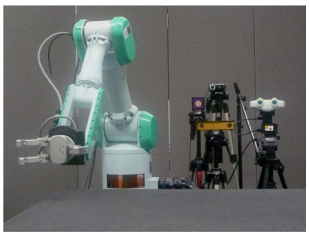


Figure 3: Robotic platform used in the experiments.

The visual features and positions of objects were extracted from image streams obtained from the stereo vision camera. The extraction and tracking of objects are done based on their color. The visual features have six dimensions: three for color (L*a*b* color space) and three for shapes. The shape features, object area f_{area}, squareness f_{sq}, and width-height ratio f_{whr} are defined as $f_{\text{area}} = wh$ and $f_{\text{sq}} = N_{\text{obj}}/wh$, where h, w and N_{obj} denotes the object's height, width, and number of pixels, respectively. For motion learning/recognition, the trajectories of objects' centers of gravity are used.

3 The LCore Framework

LCore Overview

The LCore(Iwahashi 2007) selects the optimal action based on an integrated user model trained by multimodal information when a user's utterance is input. A user model corresponding to each modality (speech, vision, etc.) is called a *belief module*. The user model integrating the five belief modules – (1) speech, (2) motion, (3) vision, (4) motion-object relationship, and (5) behavioral context– is called the *shared belief* Ψ.

Utterance Understanding in LCore

An utterance s is interpreted as a conceptual structure $z = (W_T, W_L, W_M)$, where W_T, W_L, and W_M represent the phrases describing the trajector, landmark, and motion, respectively. For motion concepts that do not require a landmark object, $z = (W_T, W_M)$. For example, the user's utterance, "*Place-on Barbabright red box,*" is interpreted as follows:

$W_T : [Barbabright], \quad W_L : [red\ ,\ box\ ], \quad W_M : [place\text{-}on]$

The LCore does not deal with function words such as prepositions and articles, i.e. the user is not supposed to use words such as "on" and "the."

Suppose that an utterance s is given under a scene O. O represents the visual features and positions of all objects in the scene. The set of possible actions A under O is defined as follows:

$$A = \{(i_t, i_r, C_V^{\langle j \rangle}) \mid i_t = 1,...,O_N,\ i_r = 1,...,R_N,\ j = 1,...,V_N\} \triangleq \{a_k \mid k = 1,2,...,|A|\}, \quad (1)$$

where i_t denotes the index of a trajector, i_r denotes the index of a reference object, O_N denotes the number of objects in O, R_N denotes the number of possible reference objects for the verb $C_V^{\langle j \rangle}$, and V_N denotes the total number of C_V in the lexicon.

Each belief module is defined as follows: First, the belief module of speech, B_S, is represented as the log probability of s conditioned by z. Here, word/phrase orders is learned by using bigrams/trigrams. Next, the belief module of motion, B_M, is defined as the log likelihood of a probabilistic model given the maximum likelihood trajectory $\hat{\mathcal{Y}}_k$ for a_k. The belief module of vision, B_I, is represented as the log likelihood of W_T given Object i's visual features $\mathbf{x}_I^{\langle i \rangle}$, where Object i is either the trajector i_t or the landmark i_r. Similar to B_I, the belief module of motion-object relationship, B_R, is represented as the log likelihood of a probabilistic model given the visual features of Objects i_t and i_r. The belief module of behavioral context, $B_H(i, \mathbf{q}^{\langle i \rangle})$, represents the adequateness of Object i as the referent under the context $\mathbf{q}^{\langle i \rangle} = (q_1^{\langle i \rangle}, q_2^{\langle i \rangle})$, where $q_1^{\langle i \rangle}$ and $q_2^{\langle i \rangle}$ stand for truth values representing the statements "Object i is being grasped" and "Object i was manipulated most recently", respectively. The details of the definitions of above modules are presented in (Iwahashi 2007).

The shared belief function Ψ is defined as the weighted sum of each belief module:

$$\begin{aligned}\Psi(s, a_k, O, \mathbf{q}^{\langle i_t \rangle}) = \\ \max_z \Big\{ & \gamma_1 \log P(s|z) P(z; G) && [B_S] \\ & + \gamma_2 \left(\log P(\mathbf{x}_I^{\langle i_t \rangle} | W_T) + \log P(\mathbf{x}_I^{\langle i_r \rangle} | W_L) \right) && [B_I] \\ & + \gamma_3 \log P(\hat{\mathcal{Y}}_k | \mathbf{x}_p^{\langle i_t \rangle}, \mathbf{x}_p^{\langle i_r \rangle}, C_V^{\langle j \rangle}) && [B_M] \\ & + \gamma_4 \log P(\mathbf{x}_I^{\langle i_t \rangle}, \mathbf{x}_I^{\langle i_r \rangle} | C_V^{\langle j \rangle}) && [B_R] \\ & + \gamma_5 \left(B_H(i_t, \mathbf{q}^{\langle i_t \rangle}) + B_H(i_r, \mathbf{q}^{\langle i_r \rangle}) \right) \Big\}, && [B_H] \end{aligned} \quad (2)$$

where $\mathbf{x}_p^{\langle i_r \rangle}$ denotes the position of Object i, and $\gamma = (\gamma_1, ..., \gamma_5)$ denotes the weights of the belief modules. The MCE learning(Katagiri, Juang, and Lee 1998) is used for the learning of γ.

Inappropriate speech recognition results are re-ranked lower by using Ψ. There are several methods for re-ranking an utterance hypothesis (e.g. (Lemon and Konstas 2009)). In contrast, information on physical properties such as vision and motion is used in Ψ, since object manipulation requires physical interaction.

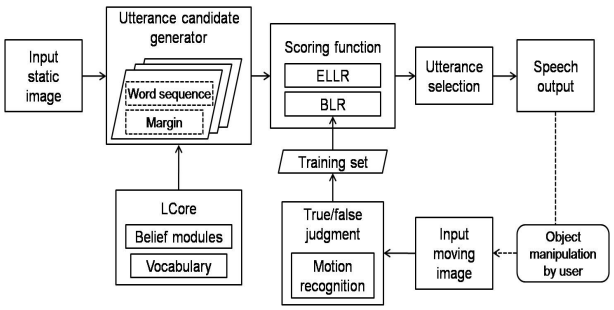


Figure 4: Schematic of the proposed method.

4 Active Learning of the Integrated Confidence Measure Function

The schematic of the proposed method is illustrated in Figure 4. Each function in the figure is explained below.

Modeling Confidence for Utterance Understanding

The proposed method quantifies ambiguities in a user's utterances. In this subsection, we first explain the ambiguity criterion used in this study.

Given a context $\mathbf{q}$, a scene O, and an utterance s, the optimal action $\hat{a}_k$ is obtained by maximizing the shared belief function.

$$\hat{a}_k = \underset{a_k \in A}{\operatorname{argmax}} \Psi(s, a_k, O, \mathbf{q}) \quad (3)$$

We define the *margin* function d for the action $a_k \in A$ as the difference in the Ψ values between a_k and the action maximizing Ψ, $a_j \ \ (j \neq k)$:

$$d(s, a_k, O, \mathbf{q}) = \Psi(s, a_k, O, \mathbf{q}) - \max_{j \neq k} \Psi(s, a_j, O, \mathbf{q}) \quad (4)$$

Let a_l be an action that gives the second maximum Ψ value. When the margin for the optimal action $\hat{a}_k$ is almost zero, the shared belief values of $\hat{a}_k$ and a_l is nearly equal; this means that the utterance s is a likely expression for both $\hat{a}_k$ and a_l. In contrast, a large margin means that s is an unambiguous expression for $\hat{a}_k$. Therefore, the margin function can be used as a measure of the utterance's ambiguity.

Now we define the *integrated confidence measure* (ICM) function by using a sigmoid function, as follows:

$$f(d; \mathbf{w}) = \frac{1}{1 + \exp^{-(w_1 d + w_0)}}, \quad (5)$$

where d is the value of the margin function for an action, and $\mathbf{w} = (w_0, w_1)$ is the parameter vector. The ICM function is used for modeling the probability of success.

We now consider the problem of estimating the parameters $\mathbf{w}$ of the ICM function based on logistic regression. The ith training sample is given as a pair consisting of the margin d_i and teaching signal u_i. Thus, the training set $\mathbb{T}^{\langle N \rangle}$ contains N samples:

$$\mathbb{T}^{\langle N \rangle} = \{(d_i, u_i) | i = 1, ..., N\}, \quad (6)$$

where u_i is 0 (failure) or 1 (success).

BLR(Genkin, Lewis, and Madigan 2007) is used for obtaining the MAP estimate of $\mathbf{w}$. A univariate Gaussian prior with mean m_i and variance τ_i $(i = 0, 1)$ on each parameter

w_i is used.

$$P(w_i|\tau_i) = \mathcal{N}(m_i, \tau_i) = \frac{1}{\sqrt{2\pi\tau_i}} \exp \frac{-w_i^2}{2\tau_i} \quad (7)$$

Utterance Selection as Active Learning

The utterance candidate generator (see Figure 4) generates all possible linguistic expressions for each action a_k and calculate their margin. The set of margin is input to the scoring function based on Expected Log Loss Reduction (ELLR)(Roy and McCallum 2001) and Bayesian logistic regression (BLR)(Genkin, Lewis, and Madigan 2007). Next, the optimal margin linked with an utterance is selected and the utterance is output as a speech command to the user. The true/false judgment module recognizes the motion performed by the user, and judges the result as 0(false) or 1(true). The result is input to the training set and used by the scoring function.

Basically, a training sample for learning the ICM function is obtained when a robot has executed a motion. Note that we assume that belief modules and Ψ are shared by the user and the robot to introduce active learning. Based on this assumption, we can train the ICM function by using training data obtained when the robot commands the user by speech to manipulate an object.

The proposed method selects the utterance that is most effective for learning the function based on Expected Log Loss Reduction (ELLR)(Roy and McCallum 2001). Among many criteria, uncertainty sampling(Lewis and Gale 1994) is the most basic method in active learning, however it selects a sample with the most entropic prediction. In contrast, ELLR asks for labels on examples that, once incorporated into training, will result in the lowest expected error on the test set(Roy and McCallum 2001).

Now, let $\hat{f}^{\langle N \rangle}(d)$ denote the ICM function trained by the data set $\mathbb{T}^{\langle N \rangle}$. The log loss $L(\mathbb{T}^{\langle N \rangle})$ is defined as follows:

$$L(\mathbb{T}^{\langle N \rangle}) = \sum_{i=1}^{N} \left\{ \hat{f}^{\langle N \rangle}(d_i) \log \hat{f}^{\langle N \rangle}(d_i) + (1 - \hat{f}^{\langle N \rangle}(d_i)) \log(1 - \hat{f}^{\langle N \rangle}(d_i)) \right\}$$

In this case, $L(\mathbb{T}^{\langle N \rangle})$ can be regarded as the sum of entropy.

Let $V = \{v_j | j = 1, ..., |V|\}$ denote the utterance candidates in the scene O, and e_j denote the margin linked with v_j. Here, V means the possible combinations of a word sequence that consists of learned words. We make V a finite set by limiting the length of a sequence. The proposed method selects the utterance that minimizes the Expected Log Loss $E(\mathbb{T}^{\langle N \rangle}, e_j)$. $E(\mathbb{T}^{\langle N \rangle}, e_j)$ is defined as follows:

$$E(\mathbb{T}^{\langle N \rangle}, e_j) = \hat{f}^{\langle N \rangle}(e_j) L(\mathbb{T}_+^{\langle N+1 \rangle}) + (1 - \hat{f}^{\langle N \rangle}(e_j)) L(\mathbb{T}_-^{\langle N+1 \rangle}),$$
$$\mathbb{T}_+^{\langle N+1 \rangle} \triangleq \mathbb{T}^{\langle N \rangle} \cup (e_j, 1), \quad \mathbb{T}_-^{\langle N+1 \rangle} \triangleq \mathbb{T}^{\langle N \rangle} \cup (e_j, 0) \quad (8)$$

Thus, Equation (8) takes into account the effect of a not-yet-obtained sample. In ELLR, $\hat{f}^{\langle N+1 \rangle}(e_j)$ is trained in advance of obtaining the $(N+1)$th sample. On the other hand, uncertainty sampling(Lewis and Gale 1994) does not take into account the effect of selecting the $(N+1)$th sample.

5 Experiments

Experimental Setup

To evaluate the proposed method, we conducted two kinds of experiments: (1) active learning of the ICM function, and (2) evaluation of the proposed method. The objective of Experiment (1) is to investigate the number of samples necessary for convergence of the learning. Experiment (2) was aimed at evaluating the effectiveness of using the result of Experiment (1) as the prior distribution. Although the effectiveness is unclear since the assumption that the user and robot share the ICM function is not always true, we will clarify the advantages of the proposed method.

In Experiment (1), the robot commanded the user by speech to manipulate objects based on the proposed method. This flow which starts from the robot's utterance and ends with the user's manipulation is called an episode. The maximum number of episodes was set to 30. The prior distribution of the parameter w_i was defined as a univariate Gaussian distribution. The parameters of the prior, or *hyperparameters*, were set as $(m_0, m_1, \tau_0, \tau_1) = (0, 1, 100, 100)$. The hyperparameters (m_0, m_1) were set as $(m_0, m_1) = (0, 1)$ so as to make the initial ICM function be the standard logistic sigmoid function. The maximum length of (W_T, W_L, W_M) were set to (3,3,1), respectively.

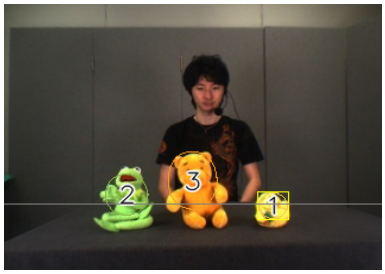
In Experiment (2), we obtained the training and test data as follows. First, the subject was instructed to command the robot in the same environment as Experiment (1). This enabled us to obtain 60 pairs of camera image and speech, which we labeled with the indices of {motion, trajector, landmark}. Half of the data was used as a training set and the other half was used as a test set.

We compared the case in which the parameter estimated in Experiment (1) was used as the prior distribution with a case involving a "standard" prior distribution without parameter tuning. The parameters of the standard prior were set as $(m_0, m_1, \tau_0, \tau_1) = (0, 1, 100, 100)$. To evaluate these methods, we compared test-set likelihood, where ten different combinations of a training and test set were used. Similar to Experiment (1), we use the word "episode" to represent the flow that begins from the user's utterance and ends with the robot's manipulation.

In Experiment (2), the number of motion failures was also compared. We compared the average number of motion failures occurring from the first to the i_cth episodes, where i_c represents the episode in which a convergence condition regarding log likelihood $\mathcal{L}$ was met. The convergence condition is set as $\mathcal{L} < -20$, based on the results of the experiment in (Sugiura et al. 2009). Although we continued the actual experiment after the convergence condition was met, the learning should be terminated here for efficiency.

The lexicon used in the experiments contained 23 words (8 nouns, 8 adjectives, and 7 verbs). The user taught the names or properties of objects in Japanese[1] by showing the objects to the robot. Unsupervised learning was used for obtaining the phoneme sequences of the words(Iwahashi 2007). Those words had been grounded to the physical prop-

[1] In this paper, the utterances are translated into English.



[Situation: Object 1 was manipulated most recently]
R: *Jump-over Pooh-doll Kermit.*
U: (The user makes Object 3 jump over Object 2.)

Figure 5: Dialogue example in the learning phase. The correct action is to make Object 3 (Pooh-doll) jump over Object 2 (Kermit).

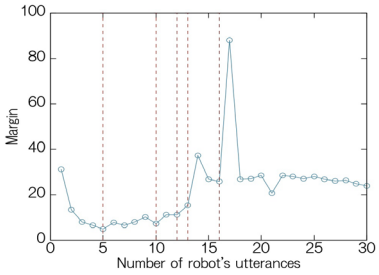


Figure 6: Margin selected by Equation (8). The dotted line shows an episode in which motion failure by the user occurred.

erties of objects and motions in the learning phase of the lexicon (Iwahashi 2007, Sugiura and Iwahashi 2007).

Results (1): Active Learning of the ICM Function

First, we address the qualitative results. Figure 5 shows an example dialogue between the subject (U) and the robot (R). In this case, the number of possible combinations of objects and motion were 45, which means that 45 pairs of a word sequence and margin are generated by the utterance candidate generator shown in Figure 4. Among the pairs, the margin $d = 13.4$ is selected based on Equation (8). Here, the utterance linked with the margin $d = 13.4$ was "*Jump-over Pooh-doll Kermit* (Make the Pooh doll jump over Kermit.)."

In Figure 6, the selected margin is plotted against an episode that represents the number of robot utterances. In the figure, the dotted line shows the episode in which motion failure by the user has occurred. From the figure, we can see that the larger margin is selected at the $(i^* + 1)$th episode compared with the i^*th episode, where i^* represents the episode in which such motion failure occurred. This means that an utterance with less ambiguity is selected at the $(i^* + 1)$th episode.

Results (2): Evaluation of the Proposed Method

Figure 7 shows examples of camera images input for the proposed method. The inputs into the system were the visual

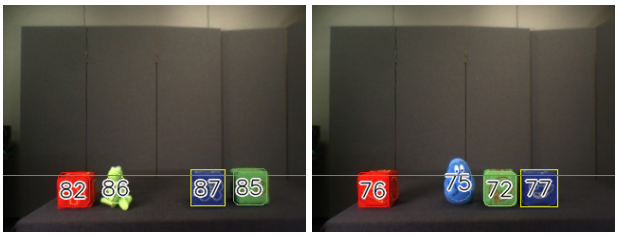


Figure 7: Training samples. Yellow frames represent the most recently manipulated objects. Left: The input user utterance was "*Move-away* Kermit," and the correct output was to move Object 86 away from Object 82. Right: The input user utterance was "*Place-on* red box," and the correct output was to place Object 76 on Object 77.

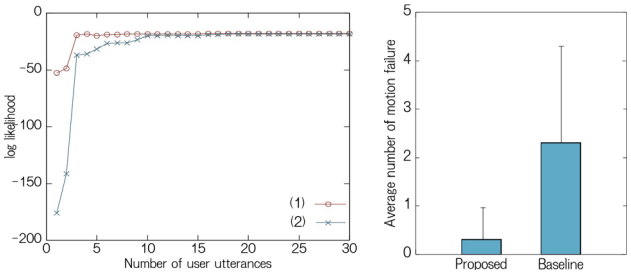


Figure 8: Left: The test-set log likelihood of (1) the proposed method and (2) a baseline. Right: Average number of motion failures.

features of the extracted objects, context information (Object X was manipulated most recently, etc), and user utterances.

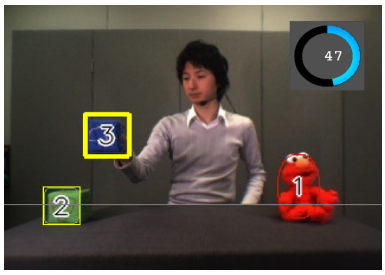
The left-hand figure of Figure 8 compares the average test-set log likelihood of (1) the proposed method and (2) the baseline which used a standard prior. The lines show the average log likelihood, where ten different combinations of a training and test set were used. The figure clearly indicates that the proposed method outperformed the baseline in the early episodes.

The right-hand figure of Figure 8 compares the average number of motion failures occurring from the first to the i_cth episodes. From the figure, we can see that the number of motion failure could be reduced by using the proposed method if we terminated the learning at the i_cth episode. The reduction in the number of motion failures supports the validity of the prior pre-trained by using active learning.

Additional Result in Motion Execution Phase

In this subsection, we show an additional qualitative result in the motion execution phase. The experiment conditions are explained in detail in (Sugiura et al. 2009). Figure 9 shows a dialogue example in which a user's utterance is disambiguated by using grounded information.

In Figure 9, the ICM value of the optimal action $\hat{a}$ was small. Therefore, the robot first asked whether "*green box*" was the trajector. Here, the word "*green*" was used to describe the major difference between Object 2 (the green box)



[Situation: Object 2 was manipulated most recently]
U: *Move-closer box Elmo.*
R: *Move-closer green box?*
U: No.
R: *Move-closer blue box?*
U: Yes.
R: (The robot moves Object 3 closer to Object 1.)

Figure 9: Dialogue example (2). Motion execution with a confirmation utterance. The correct action is to move Object 3 (the blue box) closer to Object 1 (Elmo).

and Object 3 (the blue box). In the second confirmation utterance, the word "*blue*" was inserted into the phrase W_T, since this gave the maximum margin. In contrast, the landmark was not mentioned in either generated utterance since no word insertion into W_L had a significant influence on the ICM values.

6 Conclusion

Safe interaction with users is a critically important requirement for assistive robots supporting users in everyday environments. In this paper, we proposed a method that decreases the risk of motion failure in the learning phase. One of the contributions of this study is the introduction of active learning into a multimodal spoken dialogue system. Some demo video clips can be found at `http://mastarpj.nict.go.jp/~ksugiura/video_gallery/video_gallery_en.html`.

References

Bohus, D., and Rudnicky, A. 2005. Sorry, I didn't catch that!-an investigation of non-understanding errors and recovery strategies. In *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue*.

Dale, R., and Reiter, E. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science* 19(2):233–263.

Funakoshi, K.; Spanger, P.; Nakano, M.; and Tokunaga, T. 2009. A probabilistic model of referring expressions for complex objects. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 191–194.

Genkin, A.; Lewis, D.; and Madigan, D. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics* 49(3):291–304.

Inamura, T.; Toshima, I.; Tanie, H.; and Nakamura, Y. 2004. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research* 23(4):363–377.

Iwahashi, N. 2007. Robots that learn language: Developmental approach to human-machine conversations. In Sanker, N., et al., eds., *Human-Robot Interaction*. I-Tech Education and Publishing. 95–118.

Jordan, P., and Walker, M. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research* 24(1):157–194.

Katagiri, S.; Juang, B.; and Lee, C. 1998. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. *Proceedings of the IEEE* 86(11):2345–2373.

Komatani, K., and Kawahara, T. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proceedings of the 18th conference on Computational Linguistics*, 467–473.

Krüger, V.; Kragic, D.; Ude, A.; and Geib, C. 2007. The meaning of action: a review on action recognition and mapping. *Advanced Robotics* 21(13):1473–1501.

Lemon, O., and Konstas, I. 2009. User simulations for context-sensitive speech recognition in spoken dialogue systems. In *Proceedings of EACL 2009*, 505–513.

Lewis, D., and Gale, W. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12.

Lison, P., and Kruijff, G. 2008. Salience-driven contextual priming of speech recognition for human-robot interaction. In *Proceedings of the 18th European Conference on Artificial Intelligence*.

Ogata, T.; Murase, M.; Tani, J.; Komatani, K.; and Okuno, H. G. 2007. Two-way translation of compound sentences and arm motions by recurrent neural networks. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and System*, 1858–1863.

Roy, N., and McCallum, A. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of 18th International Conference on Machine Learning*, 441–448.

Sugita, Y., and Tani, J. 2005. Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior* 13(1):33–52.

Sugiura, K., and Iwahashi, N. 2007. Learning object-manipulation verbs for human-robot communication. In *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*, 32–38.

Sugiura, K.; Iwahashi, N.; Kashioka, H.; and Nakamura, S. 2009. Bayesian learning of confidence measure function for generation of utterances and motions in object manipulation dialogue task. In *Proceedings of Interspeech 2009*, 2483–2486.

Takano, W., and Nakamura, Y. 2009. Statistically integrated semiotics that enables mutual inference between linguistic and behavioral symbols for humanoid robots. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, 2490–2496.