INTERSPEECH 2009 BRIGHTON

Bayesian Learning of Confidence Measure Function for Generation of Utterances and Motions in Object Manipulation Dialogue Task

Komei Sugiura[1], *Naoto Iwahashi*[12], *Hideki Kashioka*[1], *Satoshi Nakamura*[12]

[1]National Institute of Information and Communications Technology
[2]Advanced Telecommunications Research Institute International
{komei.sugiura, naoto.iwahashi, hideki.kashioka, satoshi.nakamura}@nict.go.jp

Abstract

This paper proposes a method that generates motions and utterances in an object manipulation dialogue task. The proposed method integrates belief modules for speech, vision, and motions into a probabilistic framework so that a user's utterances can be understood based on multimodal information. Responses to the utterances are optimized based on an integrated confidence measure function for the integrated belief modules. Bayesian logistic regression is used for the learning of the confidence measure function. The experimental results revealed that the proposed method reduced the failure rate from 12% down to 2.6% while the rejection rate was less than 24%.
Index Terms: multimodal spoken dialogue system, robot language acquisition, confidence, Bayesian logistic regression

1. Introduction

The needs of an aging society have raised the hope of robots supporting humans in daily environments. For such assistive robots, the functional capability of natural communication with users is crucial. However, the state-of-the-art techniques have safety concerns since a user's utterances are processed with non-grounded knowledge. 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.

The goal of this study is to decrease the risk. The target task of this study is called an *object manipulation dialogue* task in which a robot manipulates objects according to a user's utterances. An example of object manipulation dialogue tasks in a home environment happens when a user tells a robot to "Put the dish in the cupboard." Solving this task is fundamental for assistive robots, but it is difficult to program beforehand. This is because many candidate objects exist in the home and the desired motion depends on elements specific to each home. Therefore, in object manipulation dialogue tasks two kinds of disambiguations are necessary: (1) the disambiguation of object reference and (2) the disambiguation of motion reference.

Much work has been done to solve the disambiguation problems (e.g. [1,2]), however no previous research has realized the disambiguation of both (1) and (2). On the other hand, we have proposed the LCore framework that enables robots to learn the capability of linguistic communication from scratch [3].

In this study, we extend LCore with a scheme of dialogue management method based upon an adaptive confidence measure. The proposed method called LCore-DEC, which generates motions and utterances in an object manipulation dialogue task, is presented. LCore-DEC has two key features:

1. Bayesian logistic regression (BLR) is used for learning a confidence measure of multimodal utterance understanding so that we can estimate the utility of the robot's responses.
2. The estimated utility is then used for decision-making on the responses as motions or confirmation utterances, and for generating confirmation utterances.

BLR presents several advantages over other methods (e.g. [4]) such as (1) predicting the probability of success as a posterior probability density function, and (2) sample efficiency.

2. The LCore Method

The LCore method [3] 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.

2.1. Object Manipulation Dialogue Task

Figure 1 shows an example of an object manipulation dialogue task. 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) [5]. The reference object can be the trajector 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.

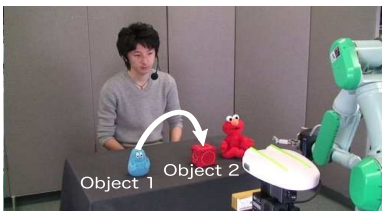


Figure 1: An example of object manipulation dialogue tasks.

2.2. 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 segments describing the trajector, landmark, and motion, respectively.

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 method 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:

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

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/segment 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_t's visual features $\mathbf{x}_I^{\langle i_t \rangle}$. 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, represents the adequateness of Object i as the referent under the context $\mathbf{q}^{\langle i \rangle}$ such as "Object i is being grasped".

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

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

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 Minimum Classification Error (MCE) learning 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. [6]). In contrast, information on physical properties such as vision and motion is used in Ψ since object manipulation needs physical interactions.

3. Learning Integrated Confidence Measure for Generation of Utterances and Motions

3.1. Modeling Confidence for Utterance Understanding

The proposed method quantifies ambiguities in a user's utterances, and generates motions or utterances as responses by maximizing a utility function. 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.

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

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 $\Psi, a_j \ \ (j \neq k)$:

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

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:

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

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.

3.2. Learning ICM Function

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 the pair of the margin d_i and teaching signal u_i, $\{(d_i, u_i) | i = 1, ..., N\}$, where u_i is 0 or 1.

BLR [7] is used for obtaining the MAP estimate of $\mathbf{w}$. A univariate Gaussian prior with mean 0 and variance τ_i $(i = 0, 1)$ on each parameter w_i is used:

$$P(w_i | \tau_i) = \mathcal{N}(0, \tau_i) = \frac{1}{\sqrt{2\pi\tau_i}} \exp \frac{-w_i^2}{2\tau_i} \tag{6}$$

3.3. Decision-Making on Multimodal Responses Based on Expected Utility

Let a^* be the action that the user intended to indicate by uttering s. For safety reasons, it is undesirable for the robot to execute an incorrect action a_k $(\neq a^*)$. A confirmation request to the user before the execution of an action can prevent the robot from executing an incorrect action. The ICM function can be used as a criterion for making a decision about whether a confirmation request is needed prior to executing the optimal action $\hat{a}_k$. In the proposed method, the ICM function is also used when confirmation request utterances are generated.

Now we consider the problem of making optimal decisions on responses to the user's utterances. We assume that the response is either the execution or confirmation of an action. The optimal response is selected based upon the expected utility.

Let b_1 be a response as a motion and b_2 be a response as a confirmation utterance. The ICM function $f(d)$ models the probability that the utterance is correctly recognized under the margin d. The expected utility $\mathbb{E}[R_i]$ for a response $b_i (i = 1, 2)$ is estimated as follows:

$$\mathbb{E}[R_i] = r_{i1} f(d) + r_{i2}(1 - f(d)), \tag{7}$$

where r_{i1} and r_{i2} denote the utility for b_i in the cases of $\hat{a}_k = a^*$ and $\hat{a}_k \neq a^*$, respectively.

The equation $\mathbb{E}[R_1] = \mathbb{E}[R_2]$ has the solution $f(d) = \theta_0$ $(0 < \theta_0 < 1)$ under the condition $r_{12} < r_{22} < r_{21} < r_{11}$. Therefore, we can use θ_0 as the threshold for selecting the optimal response $\hat{b} = \operatorname{argmax}_i \mathbb{E}[R_i]$.

3.4. Generation of Confirmation Utterances

The proposed method paraphrases object descriptions to make them more appropriate for the user and situation. Therefore, a

confirmation utterance by the proposed method is not a mere speech recognition result. To paraphrase the user's utterances, words are inserted into the segments W_T and/or W_L. The words are selected from the lexicon L based on the maximization of the margin d as follows.

Let $\psi(s,a_k,O,\mathbf{q}^{\langle i_t\rangle},z)$ be the weighted sum of belief modules. The differences between ψ and Ψ are such that ψ does not contain acoustic likelihood, and ψ is not maximized with respect to z (cf. Equation (2)). We define d_z as the margin given z:

$$d_z(s,a_j,O,\mathbf{q},z) = \psi(s,a_j,O,\mathbf{q},z) - \max_{k\neq j}\psi(s,a_k,O,\mathbf{q},z)$$

Suppose that the word set $\mathbf{c}' = \{c'_m \mid m = 1,...,M\}$ is inserted into the segment W (W_T or W_L). Here W is a sequence of words: $W \triangleq c_1c_2\cdots c_{|W|}$, where $|W|$ represents the length of W. The optimal word set $\hat{\mathbf{c}}' = \{\hat{c}'_m \mid m = 1,...,M\}$ and insertion-position set $\hat{\mathbf{p}} = \{\hat{p}_m \mid m = 1,...,M\}$ are obtained as follows:

$$(\hat{\mathbf{c}}',\hat{\mathbf{p}}) = \underset{c'_m\notin W,\,\mathbf{p}}{\operatorname{argmax}} d_z(s,a_j,O,\mathbf{q},z) \tag{8}$$

Thus, we obtain the following W' after the insertion.

$$W' = c_1\cdots c_{\hat{p}_1-1}\hat{c}'_1c_{\hat{p}_1}\cdots c_{\hat{p}_2-1}\hat{c}'_2c_{\hat{p}_2}\cdots c_{|W|} \tag{9}$$

These operations are performed for W_T and/or W_L, and finally we obtain an updated conceptual structure z':

$$z' = (W'_T,W'_L,W_M). \tag{10}$$

Based on the above, the LCore-DEC algorithm is summarized as follows:

Input Let $\langle O,\mathbf{q},s\rangle$ be an input set; a scene, behavioral context, and user's utterance.

1. Generate trajectories for all items in the action candidate set A (see Equation (1)), and obtain the shared belief values $\Psi(s,a_k,O,\mathbf{q})$ for every a_k.

2. Obtain the optimal action $\hat{a}_k$ according to Equation (3). If $f(d(s,\hat{a}_k,O,\mathbf{q})) \geq \theta_0$ holds, then execute $\hat{a}_k$ and terminate. Otherwise go to **3**.

3. Initialize the confirmation target set A' as $A' = A$.

4. Let the target action $a_j = \underset{a_j\in A'}{\operatorname{argmax}} f(d(s,a_j,O,\mathbf{q}))$. Initialize the number of inserted words, $M = 0$.

5. Increment M : $M \leftarrow M+1$, and generate z' according to Equation (10).

6. If the updated margin d' satisfies $f(d') \geq \theta_0$, go to **7**. Otherwise go to **6(a)**.

6(a) If there exists any word that can be added to z', then go to **5**. Otherwise go to **9**.

7. Make a confirmation utterance on a_j. A speech is synthesized according to z'. If W'_T or W'_L has no change from the original W_T or W_L, it is not included in the utterance.

8. If the user's response is positive, execute a_j and terminate. Otherwise remove a_j from A' and go to **8(a)**.

8(a) If A' is empty, go to **9**. Otherwise, go to **4**.

9. Reject s by uttering "Sorry, I cannot understand.", and then terminate.

4. Experiments

4.1. Experimental Setup

We conducted experiments using a platform consisting of a manipulator with seven degrees of freedom (DOFs), a four-DOF multifingered grasper, a microphone/speaker, a stereo vision camera, and a gaze-expression unit. The visual features and positions of objects were extracted from image streams obtained from the stereo vision camera. The visual features had six dimensions: three for color (L*a*b*) and three for shapes.

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 by showing the objects to the robot. Unsupervised learning was used for obtaining the phoneme sequences of the words [3]. Those words had been grounded to the physical properties of objects and motions in the learning phase of the lexicon [3, 5].

To evaluate the proposed method, we conducted two kinds of experiments: (1) learning of the ICM function, and (2) generation of utterances and motions. 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 decrease in the failure rate for the motions.

In Experiment (1), we obtained the training and test data as follows. A subject was told to sit across the table from the robot, as shown in Figure 1, and make utterances in Japanese[1] to make the robot manipulate objects. This flow which starts from the subject's utterance and ends with the robot's manipulation is called an episode. Thus, 100 pairs of utterance s and scene O were obtained. Each pair was labeled with the action intended by the subject, a^*. The average chance performance for all of the data was 2.34%, and the average number of words contained in each utterance was 2.54. Half of the data was used as a training set and the other half was used as a test set. The hyper-parameter τ_i $(i = 0, 1)$ was set to 100.

In Experiment (2), a subject had object manipulation dialogues with the robot. The training and test set were obtained in the same manner as Experiment (1). The parameters of the ICM function were trained by the training set (50 samples) and fixed in Experiment (2). The dialogue was conducted as follows. First, a sample was drawn from the test set (50 samples), and the scene O was reconstructed. Then, the recorded utterance s for the sample was input to the system, and a response was selected using the proposed method. If the response was a confirmation utterance, the user made a positive or negative response. An executed motion a_k was compared with a^* to determine whether it was correct. An episode ended if the robot executed a motion or the utterance was rejected. In Experiment (2), θ_0 was set to 0.7.

4.2. Results (1): Learning ICM Function

The qualitative results for the learning of the ICM function are shown in the left-hand side of Figure 2.

The right-hand side of Figure 2 shows a quantitative evaluation of the logistic regression. In this figure, the log likelihood $\mathcal{L}$ of the ICM function given a test set is plotted against the number of training samples. The line shows the average log likelihood, where ten different combinations of a training and test set were used. The left- and right-hand side figures reveal that $\mathcal{L}$ converged after 20 training samples. Thus, Figure 2 clearly indicates that the proposed method could give an appropriate estimation of the probability.

4.3. Results (2): Generation of Motion and Utterances

First, we address the qualitative results. Figure 3 shows an example dialogue between the subject (U) and the robot (R). The

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

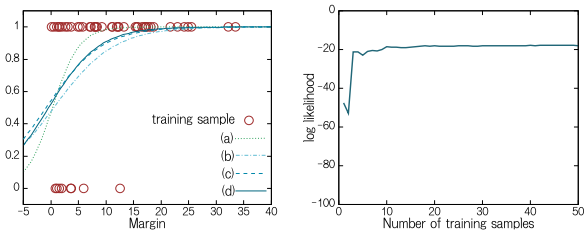


Figure 2: Left: The forms of $f(d;\mathbf{w})$ trained by (a) 10, (b) 20, (c) 30, and (d) 50 samples. Right: Average test-set log likelihood of the ICM function.

ICM value is displayed in the circle at the top right.

In Figure 3, the ICM value of the optimal action $\hat{a}_k$ was small: $f(d) = 0.478 < \theta_0$. Hence, a confirmation utterance was the optimal response. 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) and Object 3 (the blue box). In the second confirmation utterance, the word "*blue*" was inserted into the segment W_T, since this gave the maximum margin. In contrast, the landmark was not mentioned in either generated utterance since no word insertion to W_L had a large influence on the ICM values.

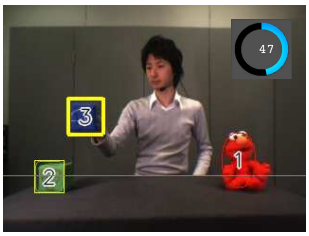
Table 1 summarizes the quantitative results of decision-making based upon the ICM values. In the table, P_f, P_r, P_c, and T_c represent the incorrect motion execution rate, rejection rate, confirmation utterance rate, and average number of confirmation requests, respectively:

$$P_f = N_f/(N_s+N_f),$$
$$P_r = N_r/N_a,$$
$$P_c = N_c/N_a,$$

where N_a, N_s, N_f, N_c, and N_r denote the number of all episodes, episodes in which correct motions were executed, episodes in which incorrect motions were executed, episodes in which confirmation utterances were generated, and episodes in which the subject's utterances were rejected (i.e. no motions were executed), respectively. Here, $N_a = N_s + N_f + N_r$. T_c means the length of interactions; for example, there were two confirmation requests in Figure 3.

Under the condition $\theta_0 = 0$, the robot always executes a motion as a response to the subject's utterance. We can regard this condition as the baseline in which the proposed method was not used. P_f was 12% (6/50) under this condition. Table 1 shows that P_f was less than 12% under other conditions, where the proposed method was used. From Table 1, we can see that P_f decreased with an increase in θ_0. Specifically, we obtained $P_f = 6.5\%$ $(3/46)$ when $\theta_0 = 0.9$ and $P_f = 2.6\%$ $(1/38)$ when $\theta_0 = 0.999$. Table 1 reveals that confirmation utterances were generated in at most half of the scenes since P_c was less than 50% in all cases. Table 1 shows that T_c was approximately 1.2 under all conditions other than those where $\theta_0 = 0$.

Finally, we investigate the rejection rate. Table 1 exhibits that P_r increased with an increase in θ_0. The episodes that ended with rejection can be categorized into two groups: (1) utterances giving $f(d) \geq \theta_0$ could not be generated for the scenes, and (2) the subject could not understand the generated utterances. An example of (1) was a scene in which no combination of the learned words could identify the trajector and/or landmark. Specifically, one of identical green boxes could not be identified since words for spacial relationships such as "right" or "below" was not learned in the experiments. An example of (2) occurred when the generated utterance included the name of an object that did not exist in the scene due to uncertainties in image processing.



[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 3: Dialogue example: Motion execution with confirmation utterances. The correct action is to move Object 3 (blue box) closer to Object 1 (Elmo).

Table 1: Evaluation of decision-making based on the ICM value

θ_0	0	0.7	0.9	0.99	0.999
P_f[%]	12.0	10.4	6.5	7.1	2.6
P_r[%]	0	4.0	8.0	16.0	24.0
P_c[%]	0	12.0	22.0	28.0	48.0
T_c	-	1.17	1.27	1.21	1.25

5. Conclusion

In this paper, we proposed LCore-DEC that generates motions and utterances in an object manipulation dialogue task to decrease the risk of incorrect motion executions by robots. One of the contributions of this study is that we integrated learning techniques studied in different research fields within a probabilistic framework; the learning of motions has been mainly studied in the robotics community, while the learning of objects has been studied in the computer vision and artificial intelligence communities. Another contribution is the introduction of utility-based dialogue management using Bayesian logistic regression into a multimodal spoken dialogue system.

Acknowledgements

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

6. References

[1] D. Roy, "Learning visually grounded words and syntax for a scene description task," *Computer Speech and Language*, vol. 16, no. 3, pp. 353–385, 2002.

[2] Y. Yamakata *et al.*, "Belief network based disambiguation of object reference in spoken dialogue system for robot," in *Proc. of the 7th ICSLP*, 2002.

[3] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Human-Robot Interaction*, 2007, pp. 95–118.

[4] D. Bohus *et al.*, "Online supervised learning of non-understanding recovery policies," in *Proc. of the IEEE/ACL Workshop on Spoken Language Technology*, 2006, pp. 170–173.

[5] K. Sugiura *et al.*, "Learning object-manipulation verbs for human-robot communication," in *Proc. of IWMISI*, 2007, pp. 32–38.

[6] O. Lemon and I. Konstas, "User simulations for context-sensitive speech recognition in spoken dialogue systems," in *Proc. of EACL*, 2009, pp. 505–513.

[7] A. Genkin *et al.*, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.