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

Komei Sugiura¹, Naoto Iwahashi¹², Hideki Kashioka¹, Satoshi Nakamura¹²

¹National Institute of Information and Communications Technology
²Advanced 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 z is interpreted as a conceptual structure z = (W_r, W_L, W_M), where W_r, 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:

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

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:

\[
A = \{ \{ i_t, i_r, C_V^{(j)} \} \mid i_t = 1, \ldots, N_t, \quad i_r = 1, \ldots, R_N, \quad j = 1, \ldots, V_N \} = \{ a_k \mid k = 1,2, \ldots, |A| \},
\]

where \( i_t \) denotes the index of a trajectory, \( i_r \) denotes the index of a reference object, \( O_{k} \) denotes the number of objects in \( O \), \( R_N \) denotes the number of possible reference objects for the verb \( C_V^{(j)} \), 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 are 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 \( \tilde{Y}_{s}^{k} \) for \( a_{k} \). The belief module of vision, \( B_{v} \), is represented as the log likelihood of a probabilistic model given the visual features of Object \( i_{t} \)’s visual features \( x_{i_{t}}^{(k)} \). Similar to \( B_{v} \), the belief module of motion-object relationship, \( B_{m} \), 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 \( i_{t} \) as the referent under the context \( q^{(k)} \) such as “Object \( i_{t} \) is being grasped”.

The shared belief function \( \Psi \) is defined as the weighted sum of each belief module:

\[
\Psi(s, a_{k}, O, q^{(k)}) = \max \left\{ \gamma_{r} \log P(s_{r}) + \gamma_{s} \log P(\tilde{Y}_{s}^{k}) + \gamma_{v} \log P(x_{i_{t}}^{(k)} | W_{T}) + \log P(x_{i_{t}}^{(k)} | W_{L}) \right. \\
\left. + \gamma_{d} \log P(x_{i_{t}}^{(k)} | W_{L}) + \gamma_{s} \left( \bar{B}_{h}(i_{t}, q_{i_{t}}^{(k)}) + \bar{B}_{h}(i_{r}, q_{i_{r}}^{(k)}) \right) \right\},
\]

where \( x_{i_{t}}^{(k)} \) denotes the position of Object \( i_{t} \), and \( \gamma = (\gamma_{r}, \gamma_{s}, \gamma_{v}, \gamma_{d}) \) denotes the weights of the belief modules. The Minimum Classification Error (MCE) learning is used for the learning of \( \gamma \).

Inappropriate speech recognition results are re-ranked lower by using \( \Psi \). 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 \( \Psi \) 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 \( q \), a scene \( O \), and an utterance \( s \), the optimal action \( a_{k} \) is obtained by maximizing the shared belief function.

\[
a_{k} = \arg \max_{a_{k} \in A} \Psi(s, a_{k}, O, q)
\]

We define the margin function \( d \) for the action \( a_{k} \in A \) as the difference in the \( \Psi \) values between \( a_{k} \) and the action maximizing \( \Psi \), \( a_{j} \) (\( j \neq k \)):

\[
d(s, a_{k}, O, q) = \Psi(s, a_{k}, O, q) - \max_{j \neq k} \Psi(s, a_{j}, O, q) \tag{4}
\]

Let \( a_{k} \) be an action that gives the second maximum \( \Psi \) value. When the margin for the optimal action \( a_{k} \) is almost zero, the shared belief values of \( a_{k} \) and \( a_{j} \) are nearly equal; this means that the utterance \( s \) is a likely expression for both \( a_{k} \) and \( a_{j} \). In contrast, a large margin means that \( s \) is an unambiguous expression for \( a_{k} \). 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:

\[
f(d; w) = \frac{1}{1 + \exp^{-w_{d}d}}, \tag{5}
\]

where \( d \) is the value of the margin function for an action, and \( w = (w_{0}, w_{1}) \) 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 \( w \) of the ICM function based on logistic regression. The \( i \)th training sample is given as the pair of the margin \( d_{i} \) and teaching signal \( \tau_{i} \), \( \tau_{i} = \{0, 1\} \), where \( \tau_{i} = 0 \) or 1.

B LR [7] is used for obtaining the MAP estimate of \( w \). A univariate Gaussian prior with mean 0 and variance \( \tau_{i} \) on each parameter \( w_{i} \) is used:

\[
P(w_{i}|\tau_{i}) = \mathcal{N}(0, \tau_{i}) = \frac{1}{\sqrt{2\pi\tau_{i}}} \exp\left(-\frac{w_{i}^2}{2\tau_{i}}\right) \tag{6}
\]

#### 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 \( a_{k} \). 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 \( E[R_{i}] \) for a response \( b_{i}(i = 1, 2) \) is estimated as follows:

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

where \( r_{i1} \) and \( r_{i2} \) denote the utility for \( b_{i} \) in the cases of \( a_{k} = a^{*} \) and \( a_{k} \neq a^{*} \), respectively.

The equation \( E[R_{1}] = E[R_{2}] \) 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 \( \theta_{0} \) as the threshold for selecting the optimal response \( b = \arg\max_{b} E[R_{i}] \).

#### 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 \( \psi(s, a_k, O, q, z) \) be the weighted sum of belief modules. The differences between \( \psi \) and \( \Psi \) are such that \( \psi \) does not contain acoustic likelihood, and \( \psi \) is not maximized with respect to \( z \) (cf. Equation (2)).

Suppose that the word set \( \mathcal{C} = \{ c_m^l \mid m = 1, \ldots, M \} \) is inserted into the segment \( W \) \((W_T \) or \( W_L)\). Here \( W \) is a sequence of words: \( W \triangleq c_1 c_2 \cdots c_{|W|} \), where \( |W| \) represents the length of \( W \).

The optimal word set \( \mathcal{C}' = \{ c_m^l \mid m = 1, \ldots, M \} \) and insertion-position set \( \mathcal{P} = \{ p_m \mid m = 1, \ldots, M \} \) are obtained as follows:

\[
\mathcal{C}', \mathcal{P} = \arg \max_{\mathcal{C}', \mathcal{P}} d_c(s, a_j, O, q, z) \tag{8}
\]

Thus, we obtain the following \( W' \) after the insertion.

\[
W' = c_1 \cdots c_{p_1} c_{l'_1} c'_{l_1} c_{p_2} \cdots c_{l'_2} c_{p_2} \cdots c_{l'_M} c_{p_M} \tag{9}
\]

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

\[
\psi' = \left( W'_T, W'_L, W_q \right). \tag{10}
\]

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

**Input**

Let \((O, q, s)\) 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 \( \Psi(s, a_k, O, q) \) for every \( a_k \).
2. Obtain the optimal action \( a_k \) according to Equation (3). If \( f(d(s, a_k, O, q)) \geq \theta_0 \) holds, then execute \( a_k \) and terminate. Otherwise go to 3.
3. Initialize the confirmation target set \( A' \) as \( A' = A \).
4. Let the target action \( a_j = \arg \max_{a_j \in A'} f(d(s, a_j, O, q)) \). Initialize the number of inserted words, \( M = 0 \).
5. Increment \( M : M \leftarrow M + 1 \), and generate \( \mathcal{C}' \) according to Equation (10).
6. If the updated margin \( d' \) satisfies \( f(d') \geq \theta_0 \), go to 7. Otherwise go to 6(a).
7. If there exists any word that can be added to \( \mathcal{C}' \), then go to 5. Otherwise go to 9.
8. Make a confirmation utterance on \( a_j \). A speech is synthesized according to \( \mathcal{C}' \). If \( W'_T \) or \( W'_L \) has no change from the original \( W_T \) or \( W_L \), it is not included in the utterance.
9. If the user’s response is positive, execute \( a_j \) and terminate. Otherwise remove \( a_j \) from \( A' \) and go to 8(a).

**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 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, \( a^\star \). 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 \( \tau_i \) \((i = 0, 1)\) 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 \( a_0 \) was compared with \( a^\star \) to determine whether it was correct. An episode ended if the robot executed a motion or the utterance was rejected. In Experiment (2), \( \theta_0 \) 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 \( L \) 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 \( L \) 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
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