2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Long-term Learning of Concept and Word by Robots: Interactive Learning Framework and Preliminary Results

Takaya Araki, Tomoaki Nakamura, and Takayuki Nagai

Abstract— **One of the biggest challenges in intelligent robotics is to build robots that can understand and use language. Such robots will be a part of our everyday life; at the same time, they can be of great help to investigate the complex mechanism of language acquisition by infants in constructive approach. To this end, we think that the practical long-term on-line concept/word learning algorithm for robots and the interactive learning framework are the key issues to be addressed. In this paper we develop a practical on-line learning algorithm that solves three remaining problems in our previous study. We also propose an interactive learning framework, in which the proposed on-line learning algorithm is embedded. The main contribution of this paper is to develop such a practical learning framework, and we test it on a real robot platform to show its potential toward the ultimate goal.**

I. Introduction

Conceptual development and language acquisition have been widely studied in the area of cognitive and developmental physiology [1]. Concepts can be thought of categories that are clustered according to perceptual similarities. Humans acquire concepts through the clustering process of everyday experiences. Words are labels that represent corresponding concepts. The meanings of words are grounded in acquired categories and the words affect the perceptual clustering process at the same time. The learning process leads to not only a problem of unsupervised clustering but also a process of interaction between the teacher and learner. In fact, infants acquire concepts and language, which enable them to understand things and to communicate with others, through the interaction between them and caregivers. It is obvious that the interaction with others is very important for infants in cognitive development [2], and it is impossible for them to acquire language without basic interaction abilities such as joint attention. This fact clearly indicates that the categorization as an unsupervised clustering scheme and the basic interaction skills are required for truly intelligent robots that can learn language over the long term.

With regard to the clustering problem, we have developed a multimodal categorization method called multimodal Latent Dirichlet Allocation （MLDA） [3], [4], which is an application of the statistical learning method in natural language processing to intelligent robotics. The MLDA has proven to be able to categorize multimodal information, i.e. audio, visual and tactile signals, in an unsupervised manner and to infer unobserved information and suitable words. In [5] the authors have extended the MLDA to an on-line version called PFoMLDA (Particle Filter on-line MLDA), that can solve the problems regarding the batch-type MLDA. Moreover, in [6], we have proposed the use of Nested Pitman-Yor Language Model （NPYLM） [7] for generating the lexicon in an unsupervised manner. Theoretically speaking, the PFoMLDA with the NPYLM makes it possible for the robot to gradually learn concepts and word meanings; the only requirement for the robot is to have the phonetic knowledge, i.e. acoustic models, for converting an input speech waveform into a sequence of phonemes.

In spite of these efforts, we still have some difficulties on the long-term learning by robots in practice. Indeed, our informal experiment gave disappointed results, which means the robot could not categorize perceptual information enough (around 40% accuracy). It turned out that there were mainly three problems in the on-line learning algorithm: 1) errors in the phoneme recognition, 2) high occurrence frequencies of functional words, and 3) batch learning of the NPYLM. In [6], we examined the impact of phoneme recognition errors on the concept learning and found that the on-line learning algorithm has tolerance to about 20% recognition error rate. However, over 20% phoneme recognition error rate drastically decreases the learning performance and this situation may easily occur in practical learning scenarios. Therefore, this paper tackles this problem in order to improve the performance of the learning algorithm. The high occurrence frequency of functional words is another problem to be solved, otherwise the robot always infers such functional words, e.g. “this”, “it”, “is” and so on, whatever the robot sees. The third problem concerns not its performance but the computational cost of the learning algorithm. In [6], the batch-type NPYLM is executed every time a new utterance is input to the system, which is an inefficient process. This paper takes a simple idea to extend the NPYLM to a pseudo on-line version, which makes the learning algorithm really efficient. One of the main issues of this paper is to solve the above problems and construct a practical on-line learning algorithm, that enables robots to learn concepts and words for long period of time.

The latter half of this paper is devoted to the interactive learning framework of the learning process. As we mentioned earlier, the interaction between the learner and teacher is a very important factor for the word learning. In order to study the relationship between language development and the

This work was supported by Grant-in-Aid for Scientific Research (C) (20500179, 23500240) and Grant-in-Aid for Scientific Research on Innovative Areas (The study on the neural dynamics for understanding communication in terms of complex hetero systems).

Takaya Araki, Tomoaki Nakamura, and Takayuki Nagai are with the Department of Mechanical Engineering and Interigent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan, {taraki, naka_t, tnagai}@apple.ee.uec.ac.jp

environment, Roy and his colleagues launched the Human Speechome Project [8]. They revealed that the caregivers change the form of their utterances in order to accommodate the linguistic knowledge of the child [8]. On the other hand, in [9], authors have proposed socially-guided learning. They analyzed human teacher's behavior, and found that their behavior affected the learning performance of a robot, and vice versa, in the context of affordance learning. These intriguing findings motivate us to pursue the dynamics of communication between human teachers and the concepts/words learning robot. To this end, we develop an interactive learning framework considering the basic skills of children. The proposed interactive learning framework, which integrates the foregoing on-line learning algorithm and interaction modules, is implemented on a real robot platform in this paper. We conduct some interactive on-line learning experiments using a few hundred objects in order to examine the possibility of lifelong learning by robots.

The ultimate goal of this study is to realize real robot intelligence and take constructive approach to elucidate the linguistic development of children. The contribution of this paper is to construct the robot system and to test its performance toward the lifelong learning.

Language acquisition by robots has been proposed in [10], [11]. In these pioneer works, they shed light on the computational model of language acquisition and showed that robots potentially be able to acquire language; however, they did not discuss long-term on-line learning. Moreover, inference among modalities have not been taken into consideration in these studies. Of course, lifelong learning by robots is drawing attention recently, e.g. [16], rather than a new idea here; there is few attempt to make robots learn concepts/words for long-term to the best of our knowledge.

Concerning categorization, image-based [12], [13], auditory-based [14], and haptic-based [15] categorizations have proposed in the literature, while this paper addresses multimodal categorization. We think that "multimodality" is one of the most important features since inference among multimodal information including words is the base of true understanding.

II. Concepts and Words Learning

A. Overview of the system and perceptual information

Figure 1 depicts an overview of the on-line learning system. The learning problem here is to estimate the parameters of the graphical model in the figure. The robot tries to find good parameters using perceptual information obtained by the robot itself and linguistic information given by human teachers. The robot platform used in this paper is shown in Fig. 2 (a). The robot consists of a 6-DOF robot arm as the base, Barrette hand, a CCD camera with kinect, a microphone, and a tactile sensor array. As shown in Fig. 2 (b), the robot autonomously acquires images of the target object form different viewpoints. A microphone mounted on the hand is used to capture the sound produced when the robot shakes the object (Fig. 2 (c)). The tactile array sensor is responsible for acquiring haptic information during the

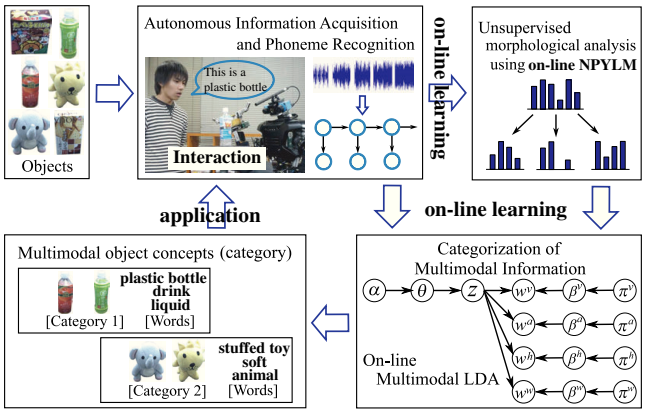

Fig. 1. Overview of the on-line learning system

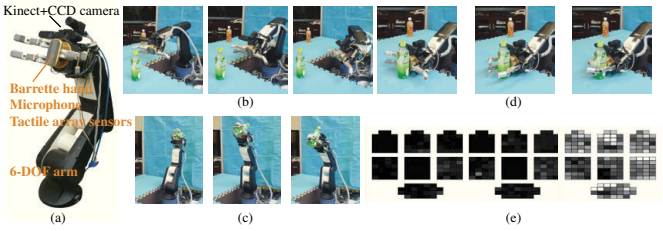

Fig. 2. Robot platform used in this study: (a) robot platform, acquisition of (b) visual information, (c) auditory information, (d) tactile information, and (e) examples of the tactile signals

grasping action (Fig. 2 (d)). When the robot pays attention to a certain object, a human teacher can describe the object including its name to the robot. The robot is supposed to have acoustic models (phoneme models) so that the input speech is transformed into a phoneme sequence. Since the robot has no lexicon, the NPYLM, which will be described later, is utilized to segment each phoneme sequence into words autonomously.

The basic idea behind the signal representation is to use the Bag of Features (BoF) model, since the BoF is successfully applied to many category recognition tasks. For visual information, 128-dimensional DSIFT (Dense Scale Invariant Feature Transform) descriptors are extracted from each image, followed by the vector quantization with a 500-dimensional code book. The MFCC (Mel Frequency Cepstral Coefficients) is used to represent the auditory signals. The MFCC vectors are vector quantized using a 50-dimensional predefined code book. The tactile information is also vector quantized resulting in a 15-dimensional histogram [5], [6]. The human utterances are segmented into words and represented by the Bag of Words (BoW).

B. Online multimodal categorization

The object concepts are represented by the MLDA as illustrated in Fig. 1. In the figure, w^v, w^a, w^h, and w^w represent visual, auditory, haptic, and word information and are assumed to be drawn from each multinomial distribution parameterized by β^v, β^a, β^h, and β^w, respectively. π^v, π^a, π^h, and π^w denote hyper parameters of Dirichlet prior distributions for β^*. z represents the category and is assumed

Algorithm 1 PFoMLDA (for a single object)

Initialize λ and α
for all m, w^m, k **do**
  $N_{mw^mk} \leftarrow (1-\lambda)N_{mw^mk}$
end for
The following process is repeated until convergence
for all m, i (of new input data) **do**
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] + (N_k^{-mi} + \alpha)\frac{N_{mw^mk}^{-mi} + \pi^m}{N_{mk}^{-mi} + W^m\pi^m}$
  end for
  $u \leftarrow$ random value $[0, 1]$
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{mi} = k$, break
    end if
  end for
end for
Select the model based on $P(w^w|\boldsymbol{w}_{obs}^v, \boldsymbol{w}_{obs}^a, \boldsymbol{w}_{obs}^h)$

to be drawn from a multinomial distribution parameterized by θ, which depends on the Dirichlet prior distribution parameterized by α. The MLDA enables the robot not only to recognize categories of unseen objects but also to infer unobserved properties of the object and words that are suitable for describing it. Inversely, understanding of the word meanings is also possible through inference using the MLDA model. We believe that this mechanism forms a base of true understanding of things and words by robots.

Originally the batch Gibbs sampling is used for the training of the MLDA. The batch algorithm relies on the assumption that the system keeps all multimodal data. Hence, a large amount of memory can be consumed as the number of teaching objects increases. Furthermore, this may take a long time and be impractical, especially in an interactive learning scenario where a human teacher is involved.

To solve this problem, the authors have proposed an on-line MLDA that sequentially updates parameters using new input data [5]. After the update of parameters, the input multimodal data can be discarded in the on-line MLDA. Basic idea behind the on-line MLDA is to use the forgetting factor λ $(0 < \lambda < 1)$ for updating the model with current parameters as initial values. In [6], we further improved the on-line MLDA using the particle filter called PFoMLDA (Particle Filter on-line MLDA). The PFoMLDA selects the forgetting factor λ and hyper parameters autonomously using the particle filter that keeps tracking good models based on the word prediction accuracy. The outline of the algorithm is given in Algorithm 1, where N_{mw^mk}, N_{mk}, N_k, and W^m are, respectively, the frequency of assigning w^m to a category k for a modality m of the object, the frequency of assigning a modality m of the object to a category k, the number of times of assigning all modalities of the object to a category k, and the dimensionality of the modality m. The superscript with the minus sign denotes exception of the data. For more details, please refer to [6].

C. Pseudo on-line NPYLM and integration with PFoMLDA

We assume that the robot has no lexicon in advance. An unsupervised word segmentation method called NPYLM [7] is utilized to segment input phoneme sequences into words. This enables the robot to acquire lexicon autonomously from scratch [6]. However the problem arises regarding the batch-type learning algorithm. As we described earlier, the batch-type algorithm has some drawbacks. The most important problem here is the computation time. In fact, the batch-type algorithm took about one hour to segment 1000 utterances in our experiment, which means that the robot cannot use taught words for a long while. Therefore it is not a good idea to use the original NPYLM in our interactive learning scenario.

Algorithm 2 Pseudo on-line NPYLM (for a single input)

a new phoneme sequence s is input:
$\mathbf{w}(s) \sim p(\mathbf{w}|s, \Theta)$
Add s to $\mathbf{S}$
Add $\mathbf{w}(s)$ to Θ
if $|\mathbf{S}| > L$ **then**
  Remove the oldest sentence from $\mathbf{S}$
end if
Blocked Gibbs sampler:
for $j \leftarrow 1$ to J **do**
  for all s in $\mathbf{S}$ **do**
    Remove $\mathbf{w}(s)$ from Θ
    $\mathbf{w}(s) \sim p(\mathbf{w}|s, \Theta)$
    Add $\mathbf{w}(s)$ to Θ
  end for
end for

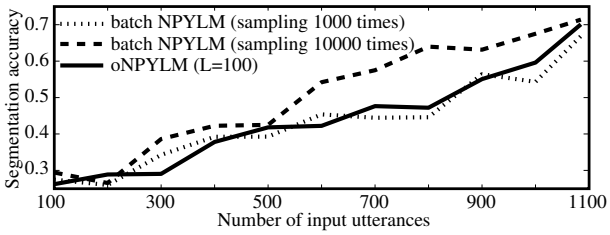

Fig. 3. Comparison of word segmentation performance

Here, we propose an on-line NPYLM (oNPYLM), which is a pseudo on-line version of the NPYLM. The idea is simple that the algorithm keeps L latest utterances and the same blocked Gibbs sampler is applied to those data with the previously converged results as the initial values. The outline of the algorithm is shown in Algorithm 2, where s, $\mathbf{S}$, $\mathbf{w}(s)$, Θ, J, and L represent, respectively, a new input utterance, a set of held utterances, the word segmentation of s, the parameters of the model, the number of iterations, and the number of utterances to be held. In our algorithm, $J = 10$ and $L = 100$ are used.

Figure 3 shows comparison of the segmentation performance between the batch-type NPYLM and the proposed oNPYLM. The solid line represents performance of the oNPYLM, while the dotted and broken lines describe performances of the NPYLM with $J = 1000$ and $J = 10000$, respectively. The oNPYLM gave the same accuracy with the NPYLM ($J = 1000$). This result is reasonable since the total number of sampling per each utterance is the same. Although the NPYLM ($J = 10000$) gave the best result

among others, the performance of the oNPYLM can be comparable if $J = 100$ is used. $J = 10$ is chosen in our algorithm because it consumes about 1 second to segment 100 phoneme sequences.

The integration of PFoMLDA and oNPYLM is straightforward; however, it should be noted that the size of the lexicon increases as the learning progresses. To deal with this, size of the BoW is set to as large as possible and an unoccupied index is assigned to a new word when it is added to the lexicon by the oNPYLM. Finally, the robot can use the histogram of words (BoW) as a part of multimodal information to do categorization.

D. Inference of unobservable information

The category of an unseen object can be inferred using the learned model. For a given model and observations regarding the novel object $\boldsymbol{w}^m_{obs}$, the most probable category $\hat{z}$ can be determined as z that maximizes $P(z|\boldsymbol{w}^m_{obs})$:

$$\hat{z} = \underset{z}{\operatorname{argmax}} \int p(z|\theta)p(\theta|\boldsymbol{w}^m_{obs})d\theta, \quad m \in \{v, a, h\}. \tag{1}$$

To recollect suitable words $\boldsymbol{w}^w$ for the unseen object, $P(\boldsymbol{w}^w|\boldsymbol{w}^m_{obs})$ is computed for given $\boldsymbol{w}^m_{obs}$ as

$$p(\boldsymbol{w}^w|\boldsymbol{w}^m_{obs}) = \int \sum_z p(\boldsymbol{w}^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^m_{obs})d\theta. \tag{2}$$

It should be noted that $p(z|\theta)$ and $p(\theta|\boldsymbol{w}^m_{obs})$ in Eqs. (1) and (2) can be updated by recalculating θ for fixed β^m using Gibbs sampling.

The problem to be solved here is that the naive implementation of Eq. (2) yields many functional words, such as "this", "is", "the", and so forth. This is because the functional words are included almost all of utterances. To overcome this problem we introduce the TF-IDF (Termed Frequency - Inverse Document Frequency) weighting scheme. Termed frequency $TF_{w^w k}$ and inverse document frequency $IDF_{w^w k}$ for a category k and a word w^w is calculated as

$$TF_{w^w k} = \frac{N_{w^w k}}{\sum_{w^w} N_{w^w k}}, \tag{3}$$

$$IDF_{w^w k} = \log \frac{K}{|\{k : k \ni w^w\}|}, \tag{4}$$

where $N_{w^w k}$, K, and $\{k : k \ni w^w\}$, respectively, represent the occurrence frequency of w^w in the category k, the number of all categories, and the number of categories containing the word w^w. The TF-IDF weight is the product of $TF_{w^w k} \times IDF_{w^w k}$ that becomes lowest when the word occurs in virtually all categories. We use this TF-IDF weight for weighting the probability $p(\boldsymbol{w}^w|\boldsymbol{w}^m_{obs})$ in Eq. (2) in order to predict category specific suitable words with higher probabilities.

E. Handling of phoneme recognition errors

As described in [6], phoneme recognition errors seriously affect the categorization result especially when the error rate reaches over 20%. To overcome this problem we use the edit distance instead of binary decision for the matching process

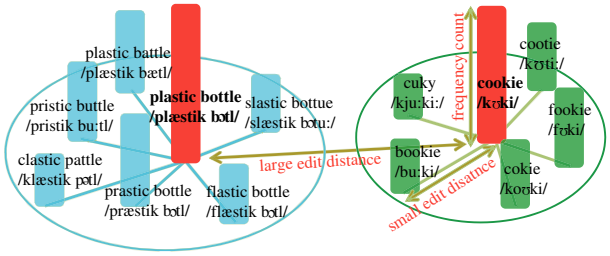

Fig. 4. Model of phoneme recognition errors

Algorithm 3 Histogram construction using edit distance

```
User's utterances s are segmented into words w
Initialize histogram h
for all w̄ in w do
    if not w̄ in the lexicon then
        Add w̄ to the lexicon
    end if
    for all w in the lexicon do
        d = EditDistance(w̄, w)
        l = max(len(w̄), len(w))
        weight = (l − d)/l
        if weight > D then
            h[w] += weight
        end if
    end for
end for
```

between two phoneme sequences. The idea is illustrated in Fig. 4. This figure depicts that the words located within a small amount of edit distance are generated from a common (true) phoneme sequence, and these words form a cluster. The edit distance between two different clusters is larger than the within class edit distance. The actual voting process to generate BoW is as follows; when a new phoneme sequence is obtained, then every word within a certain distance are voted according to the edit distance as

$$voting\ weight = \frac{(word\ length - edit\ distance)}{(word\ length)}. \tag{5}$$

Algorithm 3 shows the details of the voting algorithm. In the proposed learning algorithm, all BoWs are generated in this way. When the robot infer suitable words for the current input data, the same idea is applicable. This idea is an empirical handling of phoneme errors and there is no theoretical validity. However, this solution is plausibly reasonable and works good in our particular case as shown in the later experiments.

III. Interactive Learning Framework

As we mentioned earlier, the interactive learning framework is necessary for the word learning robots. This section proposes an interactive learning framework based on the PFoMLDA with the oNPYLM proposed in the previous section.

A. Overall architecture of the interactive learning

The overall architecture of the proposed learning framework is illustrated in Fig. 5, which consists of the on-

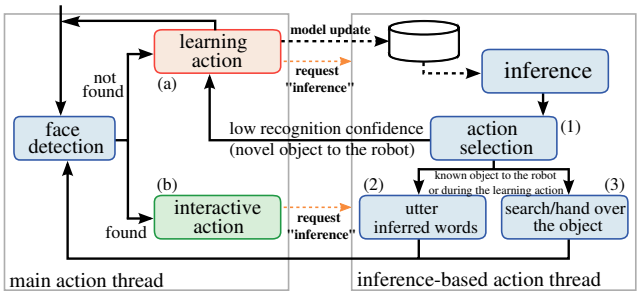

Fig. 5. Proposed architecture of interactive on-line learning framework

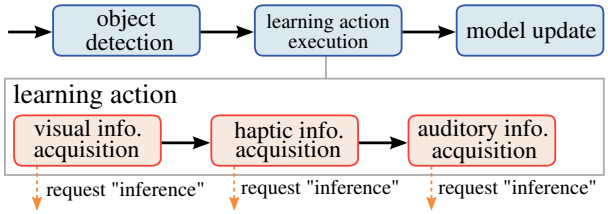

Fig. 6. Block diagram of learning action

line learning algorithm, multimodal information acquisition, and interactions with human teachers. The multithreading is involved in the proposed architecture. The left part of Fig. 5 denotes "main action thread", which manages actions of the robot. In this main thread the robot looks for a human partner at first. If the robot finds a person, then the internal state transits to the interaction part (Fig. 5 (b)). When the robot does not find any person, the robot starts finding a novel object. At this time the current model, that the robot has, is used for inferring the category of the object. This process corresponds to novelty detection. If a novel object is found, the learning process (Fig. 5 (a)), which will be described later, is activated. The learning process ends up with updating the model and then the state transits to the human detection mode.

On the other hand, the right part of Fig. 5 represents "inference-based action thread". This thread is responsible for inferring unobservable information using current input signals and the learned model at any time the request is arrived from the main thread. This thread has in total three parts: (1) action selection according to the probability of the inferred result, (2) inference of probable words form the input multimodal information, and (3) inference of visual information form the user's utterance and searching for the corresponding object.

B. Learning action

The block diagram of the learning action is shown in Fig. 6. The learning action begins with novel object detection, followed by the action sequence of multimodal information acquisition. During the execution of the series of actions, user's utterances, if detected, are phoneme recognized and processed by the oNPYLM to transform the phoneme sequence into BoW representation. All of these multimodal

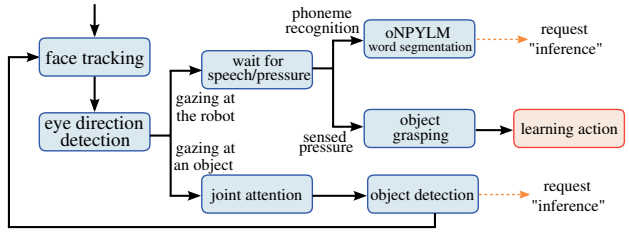

Fig. 7. Block diagram of interactive action by the robot

signals are processed and fed to the PFoMLDA to update the model. It should be noted that the oNPYLM is also updated. Moreover, the learning action includes inference of probable words from the partially input multimodal information, and utterance of these inferred words by sending a message "request inference" to the inference-based action thread. This is aimed at showing current robot's ability to the human teacher. This kind of interaction also helps for the teacher to avoid losing will to communicate with the learning robot.

C. Interactive action

The block diagram of the interactive action is given in Fig. 7. This module starts with face detection, tracking, and estimation of the gaze direction, which mimics infant's eye direction detection (EDD) [2]. When the robot detects user is gazing at the robot, it waits for speech or haptic input for a certain period of time. This is an action selection strategy to react to the user's actions such as talk to the robot, handing something, etc. If the robot detects the speech during this time, the utterance is phoneme recognized and segmented into words using the oNPYLM. Then, the robot infers visual information from the words and looks for the object that matches to the inferred visual information with enough probability. Finally, the robot grasps the object and hands it over to the human teacher. On the one hand, if the robot is aware of something in its hand, the robot grasps it and the learning action (Fig. 6) is activated.

Joint attention is also implemented on the robot. Joint attention is the shared focus of the teacher and the robot on an object. It is well known fact that joint attention plays an important role in language acquisition by infants. This motivates us to implement it on the robot. The robot can detect the gaze point using an image frame and depth information, and watch at the point to find something import to learn. If an object is found, the "inference-based action thread" is activated by signaling "request inference" so that the robot learns it if it is novel. If nothing can be found at that point, the state transits back to the face tracking mode.

The proposed framework makes the robot to learn object concepts and words through interaction with a human teacher. We believe that the most important thing is that the use of the latest concept model triggers interaction between the robot and the teacher, which elicits indispensable information to update the model. This kind of loop must be a prerequisite for observing the dynamics of human teacher's behavior.

Fig. 8. Objects used in the experiment

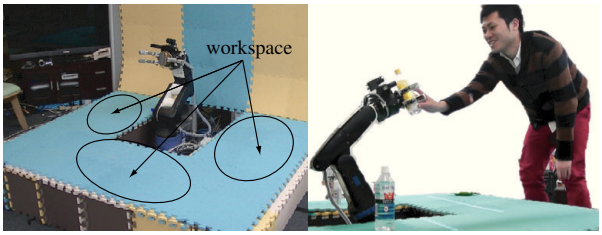

Fig. 9. Scenery of the experiment: (a) the robot and workspace, and (b) a scene of human-robot interaction

IV. Experiments

Two experiments are conducted: 1) novice users teach the robot small subset of objects to observe their interactions with the robot, and 2) the learning robot is tested for a week as a preliminary step toward long-term learning. Please note that all experiments were carried out in Japanese.

A. Experimental setup

The robot shown in Fig. 2 is used in this experiment. Figures 8 and 9 show 125 objects with 24 categories and scenery of the experiment, respectively. We define the concordance rate (accuracy) to evaluate the categorization performance as

$$Concordance = \frac{1}{Q}\sum_{j=1}^{Q}\delta(c_1(j), c_2(j)), \qquad (6)$$

where Q, $c_1(j)$, and $c_2(j)$, respectively, represent number of objects, the category index of the j-th object by the robot's categorization, and that of ground truth. $\delta(a, b)$ denotes the delta function that takes 1 if $a = b$ and 0 otherwise. It should be noted that category indexes are exchanged so that Eq. (6) is maximized.

B. Observation of interactive learning

Four test subjects are divided into two groups: half of the subjects (non-interaction condition) teach the robot 20 objects without interaction and the other half (interaction condition) freely teach the same 20 objects with the proposed interactive framework. Twenty objects with five categories are chosen from Fig. 8 . In the non-interaction condition, subjects teach the robot 20 objects in order mechanically,

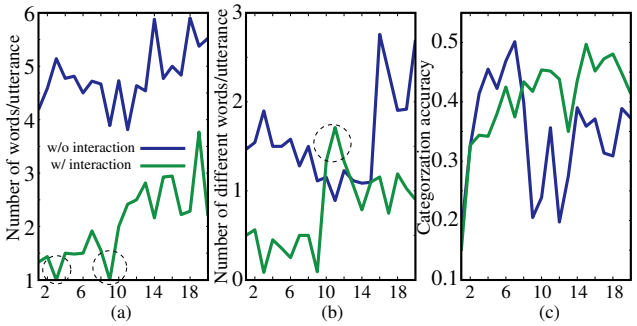

Fig. 10. Results of the interactive learning experiment: (a) MLU , (b) TTR, and (c) categorization performance of the robot

while in the interaction condition there is no restriction on the order and teaching the same object many times is allowed. The teaching process is finished after the subject taught 20 times to the robot. Therefore, in the interaction condition the subject cannot teach all of 20 objects if the same object is taught multiple times.

All utterances by the human subjects were transcribed and analyzed; the mean length of utterances (MLU) and the type-token ratio (TTR) were calculated. Figure 10 shows the results in each individual case (two out of four are give for visibility, since two subjects in the same condition showed a similar tendency). There is a big absolute difference between two groups in MLU as shown in Fig. 10 (a). Since the MLU measures complexity of sentences, it can be seen that the teacher in the interaction group started with simple and made it complex gradually. The most likely cause of this change in complexity is due to the change in ability of the robot. In fact, we found a significant correlation between the MLU and the categorization accuracy in Fig. 10 (c) only in the interaction condition ($r = .50$, $p < .05$). Interestingly, the robot's utterance of suitable words sometimes led to the deep dent in the MLU curve (marked circle in Fig. 10 (a)). For example, in the period of 9th learning, the robot said exactly correct name of the object. Before that period the teacher reduced the number of words gradually until single-word utterance. After the period, in which the robot said something correct, the teacher increased complexity of sentences. The big jump in TTR at the 9th learning (marked circle in Fig. 10 (b)) was caused by the same reason presumably. This behavior suggests short temporal dynamics between human teachers and the robot. Although the time scale is not the same, this kind of fine tuning was observed in child-directed speech [8].

C. Preliminary result on long-term learning

An on-line learning experiment was carried out using the proposed interactive learning framework. All objects shown in Fig. 8 were used multiple times and the experiment took about a week (3 to 5 hours a day). The human teacher freely selected which object to teach and the learning process was activated 200 times in total.

1) How many words did the robot actually acquire?: We first counted number of utterances by the human teacher and

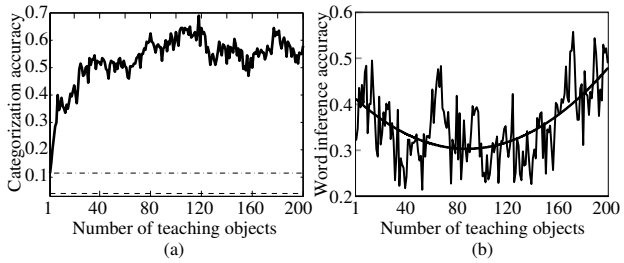

Fig. 11. Performance evaluations: (a) categorization accuracy, and (b) words inference accuracy

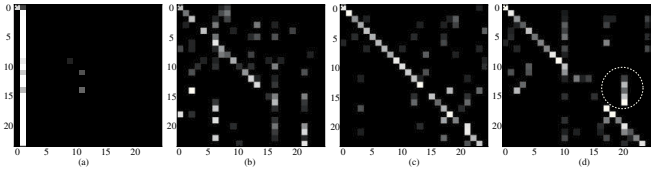

Fig. 12. Visualization of categorization results: after learned (a) 1st object, (b) 10th object, (c) 120th object, and (d) 200th object

number of acquired words by the robot so far. The number of utterances by the teacher was 1055 and the robot acquired 924 words in total. This lexicon with 924 words contains 632 meaningless words that were generated incidentally by phoneme recognition and segmentation errors. There are 58 meaningful words: 4 functional words, 10 adjectives, 40 nouns, and 4 verbs. The rest of 234 words are duplications of the 58 meaningful words with a small difference caused by the phoneme recognition errors. This result shows that only 6.3% of the whole words in the acquired lexicon (8.4 % if the duplicated words are ignored) is meaningful. This ratio seems to be quite low and indicates that special care is required in the learning phase. Our proposed voting scheme works because the 234 duplicated words are used implicitly for voting (generating Bow representation) considering the edit distance so that the frequency counts of meaningful words increase. In fact, the total weighted counts of meaningless words are about 10 % in all BoWs, which significantly improves the categorization performance.

2) Categorization performance: Performance of categorization was evaluated at each learning stage using Eq. (6). Figure 11 (a) shows the result. The categorization results are visualized in Fig. 12. From Fig. 11 (a), we can see the performance improved as the number of teaching objects increased. After 120th objects, a slow oscillation in performance is observed. This is caused by some indiscernible categories (marked with a circle in Fig. 12 (d)). For instance, shampoos and detergents are refill packs with similar appearance and softness. The bath salts, biscuits, and teabags are in the packaging boxes with similar textures and hardness as well. These categories were fluctuated over times, wheres others, such as plastic bottles, stuffed animals and so on, held relatively firm categories. The maximum accuracy of categorization was 69.0%.

3) Word inference performance: To evaluate inference performance of the models, unseen objects (each object belongs to one of 24 categories) were given to the robot and words were inferred using visual information. Each uttered word by the robot was judged whether it was suitable or not for describing the object. The accuracy of words inference over time is shown in Fig. 11 (b). The U-shaped curve in the graph can be explained as follows: In the early stages of the learning, the lexicon size was small and many functional words were included in the inferred words since the TF-IDF weighting scheme does not work in the early stages. This situation made the words inference relatively easy that led to a high inference accuracy. As the learning progress, size of the lexicon increased and the accuracy went down. Finally, further learning improved the inference accuracy.

Figure 13 illustrates comparison of inferred words for an unseen mug by the proposed method, the method without considering the edit distance (phoneme errors), and the method without using the TF-IDF. The red bar in the graph represents probability of the correct word (some phoneme errors are tolerated). The green and blue bars denote, respectively, probabilities of the incorrect words and those of the functional words. The yellow bar represents probability of the meaningless (error-related) word. From Fig. 13 (c), TF-IDF weighting scheme successfully decrease the probability of functional words. Functional words are not necessarily incorrect inferences; however it is obvious that category names and/or adjectives are preferable. Although the correct word was inferred without considering phoneme errors, inferring many similar words is the problem as can be seen in Fig. 13 (b). The proposed method solves this problem by considering the edit distance between phoneme sequences including errors (Fig. 13 (a)).

In Fig. 14 some examples of words inference are shown. Although some phoneme errors are included, correct words are selected in (a)-(f). It should be emphasized that all words were acquired from scratch using acoustic models and all inferences were carried out for unseen objects, which were not included in the training process.

On the other hand, Fig. 14 (g), (h), and (i) show false inference results. The errors in (g) and (h) are due to incorrect segmentation of the phoneme sequences. In (g) the correct phoneme sequence /garagara/ was divided into two. In (i), false category recognition caused the incorrect words inference. In fact, the box of the bath salt looks quite similar to other boxes such as biscuit, juice carton, etc.

V. Conclusion and Future Work

This paper discussed long-term learning of concepts and words by robots. We developed a practical on-line learning algorithm, and it was integrated within the proposed interactive learning framework. Preliminary experiments validated the proposed interactive framework. Although promising results were obtained, we need to examine further how much would the robot learn. Currently, the robot keeps learning and we are working on the analysis of the experimental data which grows everyday. Moreover, we have to increase the number of test subjects as the teacher to observe the dynamics between human teachers and robots.

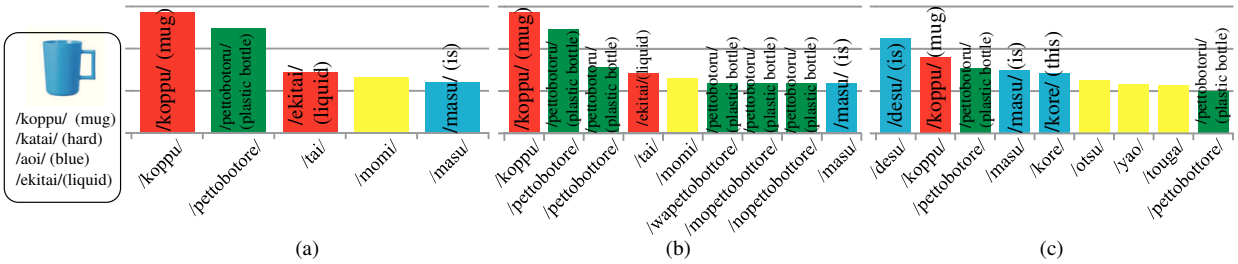

Fig. 13. Comparison of inferred words for a mug by: (a) the proposed method, (b) without considering phoneme errors, and (c) without using TF-IDF

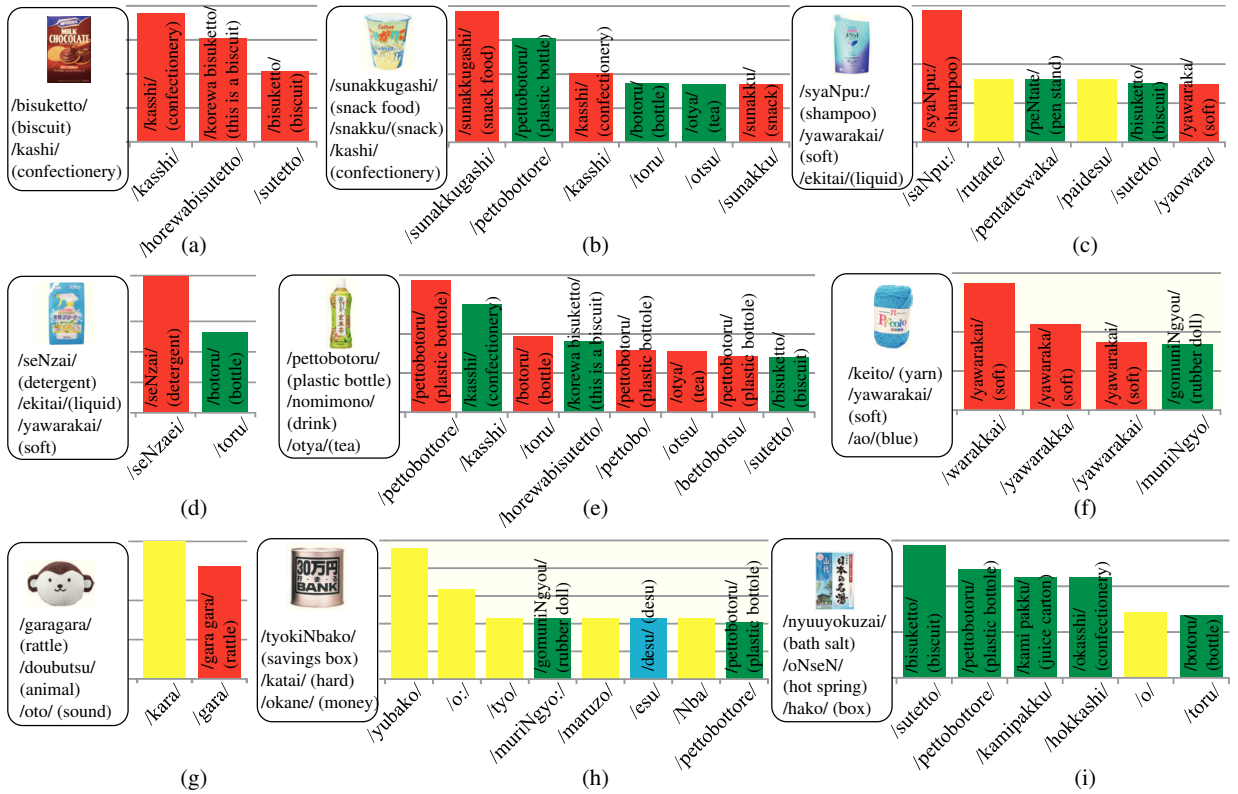

Fig. 14. Examples of inferred words: in (a)–(f) the correct word is inferred with the highest probability, while (g)-(i) incorrect word is inferred as the top

REFERENCES

[1] F. G. Ashby, and W. T. Maddox, "Human Category Learning," Annu. Rev. Psychology, 56, pp. 149–178, 2005

[2] S. B.-Cohen, Mindblindness, MIT Press, Cambridge MA, 1995

[3] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Object Categorization by a Robot," in Proc. IROS, pp. 2415–2420, 2007

[4] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of Word Meanings in Multimodal Concepts Using LDA," in Proc. IROS, pp. 3943–3948, 2009

[5] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Autonomous Acquisition of Multimodal Information for Online Object Concept Formation by Robots," in Proc. IROS, pp. 1540–1547, 2011

[6] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online Learning of Concepts and Words Using Multimodal LDA and Hierarchical Pitman-Yor Language Model," in Proc. of IROS, pp.1623–1630, 2012

[7] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling," in Proc. ACL- IJCNLP, Volume 1, pp.100–108, 2009

[8] B. C. Roy, M. C. Frank, and D. Roy, "Exploring Word Learning in a High-Density Longitudinal Corpus," in Proc. of the 31st Annual Meeting of the Cognitive Science Society, pp.2106–2111, 2009

[9] A. L. Thomaz, and M. Cakmak, "Learning about Objects with Human Teachers," in Proc. of HRI'09, pp.15–22, 2009

[10] D. Roy, and A. Pentland, "Learning Words from Sights and Sounds: A Computational Model," Cognitive Science, Vol.26, No.1, pp.113–146, 2002

[11] N. Iwahashi, Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations, In N.Sankar ed. Human-Robot Interaction, pp.95–118, I-Tech Education and Publishing, 2007

[12] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering Object Categories in Image Collections," in Proc. ICCV, pp. 370–377, 2005

[13] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," in Proc. CVPR, pp. 1903–1910, 2009

[14] J. Sinapov, and A. Stoytchev, "Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning," in Proc. ICRA, pp. 184–190, 2011

[15] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, "Object Identification with Tactile Sensors Using Bag-of-Features," in Proc. IROS, pp. 243–248, 2009

[16] R. Triebel, and L. Spinello, "Lifelong Learning for Mobile Robotics Applications," IROS 2012 Workshop, 2012