LANGUAGE ACQUISITION THROUGH A HUMAN-ROBOT INTERFACE

Naoto Iwahashi
Sony Computer Science Labs Inc.
Takanawa Muse Bldg. 3-14-13 Higashi-Gotanda Shinagawa-ku, Tokyo, 141-0022 Japan,
iwahashi@cs.sony.co.jp

ABSTRACT

This paper describes an algorithm for spoken language acquisition through a human-robot interface based on speech, vision, and behavior. In this algorithm the grounded language knowledge is represented by graphical statistical models consisting of hidden Markov models and stochastic context-free grammar. The learning of the lexicon is based on the independence between speech and visual features in each of lexical items. In the grammar-learning process, the syntactic structure of each spoken utterance is inferred from the conceptual structure extracted from the visual observation. The algorithm is robust against ambiguity and sparseness of learning data because it is based on information-theoretical learning.

1. INTRODUCTION

The recent progress of such interactive sytem technologies as computation, telecommunications, sensing, and robotics has made the development of natural-language interfaces with machines more important because it has increased the demand for easy and comfortable relationships with machines. The study of language acquisition by machines has also been attracting interest in various research areas, and the recent progress in both linguistics and machine learning has made it a very fruitful field of study [1]. The use of language acquisition schemes in interactive systems has thus been studied in attempts to increase the flexibility of the language interface, and one of the practical applications investigated was an automatic call-routing system using speech recognition [2]. In intelligent interactive systems working in real environment, the language acquisition scheme must cope with the following two major problems:

- Various aspects of spoken language are open – vocabulary, sentence expression, meaning, speaking style, speaker’s voice characteristics, and so on – and are hard to restrict even when the functions of the systems are limited.
- Spoken language should be grounded in the real world.

Several pioneering studies have developed language acquisition algorithms based on inductive learning using a set of pairs each consisting of a word sequence and either its nonlinguistic information or its semantic information. For example, [3] described a word-to-meaning mapping algorithm using a set of pairs each consisting of a sentence and a collection of its possible meanings represented symbolically with Jackendoff-style expression. It was based on cross-situational learning [4] and successfully addressed the problems due to homonyms and to noisy learning data. Visual rather than symbolic information has been used in word-to-meaning learning tasks [5, 6, 7, 8], and the judgement of whether or not the system’s response is appropriate has also been used as nonlinguistic information [9, 10]. And spoken-word acquisition algorithm based on unsupervised clustering of speech tokens, without the learning data being divided into any clusters beforehand, was presented in [2]. This algorithm used template-pattern-matching method. There have also been some studies on the use of semantic information in the learning of syntactic rules [11, 12]. An algorithm for the learning of stochastic regular grammar in a visually grounded way was presented in [6], in which the order of words in utterances was dealt with. In all these algorithms, however, some categories of the observed linguistic and nonlinguistic information, such as phonemes and meanings, were specified beforehand, and these categories were not expandable. The learning algorithms were therefore based mainly on symbolic or discrete information, although some also used continuous information. Another shortcoming of these algorithms is that they do not deal with temporal information in linguistic and non-linguistic learning data simultaneously, although spatiotemporal information should be processed in natural communication.

This paper describes an algorithm by which the grounded language knowledge can be acquired through a human-robot interface based on speech, vision, and behavior. In grammar acquisition uses the semantic bootstrapping scheme [13], where speech is associated with visual and behavioral observations through conceptual structures extracted from them. The algorithm described here is the extension of an earlier one that used as nonlinguistic learning data only graphical objects displayed on a screen [14].

2. LEARNING TASK

The spoken language acquisition task in the present work was set up as follows. A camera unit and a robot arm with a hand are set alongside a table, and a person and the learning system see and move the objects on the table as shown in Fig. 1. The system initially does not have any concepts about the specific objects and does not know any words. To teach the system, the person speaks into a microphone while pointing to and moving the objects
on the table. The person speaks slowly and pauses briefly between words. The system learns language through a sequence of such learning strokes, which provides the set of the scenes and the speech describing them. During the course of the task, the person can confirm how well the system has learned by asking the system to speak about a given scene and or by asking the system to move objects.

3. LANGUAGE ACQUISITION ALGORITHM

3.1. Outline

The system initially has no lexicon. It acquires linguistic and nonlinguistic knowledge, and the knowledge it obtains is represented by graphical statistical models consisting of normal probability density functions (p.d.fs), hidden Markov models (HMMs), stochastic context-free grammar (SCFG), and a joint probability network (Fig. 2). These statistical models are learned inductively by using the set of the episodes (the set of visual observations and the speech describing them).

The system first learns the lexicon, then grammar. In the lexicon-learning phase, when the system’s attention is directed to an object and the speech describing it, the speech and visual observations are associated. The set of the associated speech-visual observations is divided into clusters, and the membership functions represented by statistical models are learned using the observations in each cluster to form the lexical items. In the grammar-learning phase, when the system’s attention is directed to the person’s action to move an object and the speech describing it, the sequences of the words in the speech are recognized. Then possible concepts corresponding to recognized words are extracted from the scene during each action, and then a possible structural relation among them is obtained by scene analysis. This conceptual structure is the candidate meaning of the given speech, and grammar is learned by inferring the syntactic structure of the speech from the conceptual structure. The conceptual structure model is learned as the accumulation of perceptual experience and represents the information of the co-occurrence of the concepts in conceptual structures. The complexity of the statistical models grows automatically according to given learning data. The system understands speech and generates actions, by using the lexicon, the grammar, and the conceptual structure model. No text is dealt with in either the input or output of the system.

3.2. Attention control

The system directs its attention to the objects which have been put on the table, the ones which is being touched by the person, and the moving ones. Once the system directed its attention to particular objects, it is kept on them for a while unless the person directs its attention to different objects. When the attention is given to objects and the person speaks almost simultaneously, the visual observation about that objects and the speech observation are associated. The set of pairs of the associated speech and visual observations is used for learning.

3.3. Lexicon acquisition

The raw speech signal and visual information of objects are first mapped into the perceptually appropriate feature spaces. The features of speech and objects are respectively decided by the requirements derived from speech recognition and by the requirements derived from conceptual scene analysis. Speech and the movement of the objects are represented by the time sequence of these feature vectors.

The set of the speech-visual observations is the divided into two sets: one including static visual observations and the speech observations associated with them, and the other including dynamic visual observations and the speech observations associated with them. Then the observations in each of these two sets are divided into clusters. Each cluster corresponds to a lexical item, which consists of a concept membership function and a word membership function. The clustering method is described in the next subsection. Finally, the concept membership functions and the word membership functions are respectively learned using the visual observations and the speech observations in each clusters. The membership functions are represented by statistical models, and are used as discriminant functions in speech recognition and scene analysis. HMMs, each of which has left-to-right state structure, are used to represent the membership functions for the concepts of dynamic characteristics about objects and words.
The number of states in each HMM is determined by cross validation among learning samples, and the output p.d.f. at each state is represented by a multivariate normal distribution. Multivariate normal p.d.f.s are used to represent the membership functions of static concepts about objects. To reduce the severity of the problem of data sparseness and to make it easy to use high-dimensional features, the Bayesian learning method [15] is used for the learning of the multivariate normal p.d.f.s characterizing the static concepts.

3.4. Clustering for lexicon acquisition

The clustering is carried out based on maximization of the mutual information between speech $S$ and visual information $V$. The problem here is that the system initially does not know how many different words the person has spoken during the teaching of the lexicon. The system should determine the number from the observation data.

We assume that the associated speech observation and visual observation respectively are the members of the word and concept in same lexical item. We can consider that speech and visual features are independent in each of lexical items

$$p(S, V|c) = p(S|c)p(V|c) \quad \forall c \in L,$$

where $L$ and $c$ respectively denote true lexicon and true lexical item. This means that the value of the mutual information conditioned by the lexical item is 0:

$$I(S; V|c) = \int p(s, v|c) \log \frac{p(s, v|c)}{p(s|c)p(v|c)} dsdv = 0 \quad \forall c \in L. \quad (1)$$

Thus, splitting of the cluster consisting of the speech-visual observations, all of which are the members of same lexical item, does not increase the mutual information, which are not conditioned by the lexical items. Based on this, the system infers the number $m$ of lexical items as the smallest number of clusters which maximize the mutual information written as

$$I(S; V|L_m) = \int p(s, v)$$

$$\times \frac{\sum_{i=1}^{m} \{p(s, c_i^{m})p(v, c_i^{m})P(c_i^{m})\}}{\sum_{i=1}^{m} \{p(v|c_i^{m})P(c_i^{m})\}} dsdv, \quad (2)$$

where $c_i^{m}$ denotes the $i$th lexical item in the estimated lexicon $L_m$ which includes $m$ lexical items. $p(s|c_i^{m})$ and $p(v|c_i^{m})$ are respectively the word and concept membership functions. This clustering method is described in detail in [16].

3.5. Conceptual scene analysis

The system analyzes each scene to obtain its conceptual structure, which is represented by a conceptual expression consisting of some of the lexical items and their semantic attributes, such as [object], [action], and [to]. These semantic attributes are initially given to the system and are fixed. The scene analysis assigns semantic attributes to each extracted concept. For instance, if the concept lift is extracted as the previous movement of an object, the attribute [object] is assigned. And if it is extracted as the current movement of a object, which is caused by the person’s action, the attribute [action] is assigned. When the person’s operation is

**to move the red ball, which has lifted, onto the blue block on the right-hand side,**

the conceptual expression might be

$$[\text{action}] : \text{move}$$

$$[\text{object}] : \text{lift, red, ball}$$

$$[\text{to}] : \text{blue, block, right}$$

where the items in the right-hand column are concepts in the lexicon. Note that although the concepts are denoted by text here for convenience, in the algorithm they are identified by the indices assigned by the system. The conceptual expression is constructed using the individual concepts of possible words in a speech such that the likelihood of the membership function made by the composition of the individual concept membership functions is maximized for the visual observation.

3.6. Grammar learning

Grammar is learned through adaptation of the initial neural stochastic grammar, and learnable grammar is rather restricted so far: functional words such as prepositions and articles, for example, are not treated. Let a constituent of a sentence be defined as a word group that describes a concept, which could be a structural combination of multiple concepts. Each constituent is characterized by the semantic attribute assigned to the concept that the constituent describes. Let $a_i$ $(i = 1, 2, \ldots)$ be the semantic attributes. The stochastic grammar $SG$ consisting of probabilities $P(a_i \rightarrow a_j a_k)$ that the constituent with semantic attribute $a_i$ consists of two constituents each with semantic attribute $a_j$ and $a_k$ in this order and probabilities $P(a_i \rightarrow c_{j k} \cdots)$ that the constituent with semantic attribute $a_i$ consists of word sequence ‘$c_{j k} \cdots$’. The syntactic structure of each utterance is inferred using the conceptual structure extracted by the scene analysis process. The following is an example of the inference procedure. Ideally, the speech recognizer recognizes the word sequence “lift ball blue block move” perfectly. Then the scene analyzer produces a conceptual structure by using

These can be defined task-dependently. We may be able to use the semantic primitives described in [17] and [18], although the extraction of such semantic primitives from a scene is problematic.
the concepts corresponding to these words. The comparison of the recognized word sequence with the conceptual structure results in the sentence being divided into the following three constituents:

\[
\text{([object], (left, ball), (to), (blue, block), (action, move))}
\]

The SG is adapted using the information of the order ([action], [object], [to]) of the constituents’ attributes.

Because the input utterance itself could be ungrammatical or not describe the scene correctly, the algorithm utilizes Bayesian learning so that the values of SG probabilities are adapted robustly, that is, a small number of improper learning samples do not much influence the grammar adaptation.

3.7. Learning of conceptual structure model

The statistical model of the conceptual structure is represented by the probabilities of the occurrence of the concepts in the conceptual structure extracted from the scene. The probability that concept \( c_i \) and \( c_j \) respectively occur with semantic attributes \( a_k \) and \( a_l \) simultaneously is calculated by using a \textit{a posteriori} probabilities as

\[
P((c_i, a_k), (c_j, a_l)) = \frac{1}{N} \sum_{n=1}^{N} (P(c_i, a_k|o_n)P(c_j, a_l|o_n))
\]

where \( o_n \) denotes the \( n \)th speech-visual observation. The \textit{a posteriori} probabilities are calculated by using the likelihood values of possible concepts with regard to each speech-visual observation.

3.8. Speech understanding and action generation

The meaning of an utterance asking the system to move an object is inferred by using both linguistic knowledge and nonlinguistic knowledge. One of the objects on the table is accordingly selected and moved by the robot arm. The trajectory of the movement of the robot arm to move the object is calculated based on the maximum likelihood criterion with regard to dynamic concept HMMs [19], with introducing the constraint in terms of initial position of an object. If the utterance includes the dynamic concept \( c_d \) for action but does not include enough information to identify the object to be moved, the selection of an appropriate object is based on the following two principles:

- An object that has a concept \( c_t \) such that the value of \( P((c_t, \text{[object]}), (c_d, \text{[action]}) \) is high should be selected.
- The object whose movement would be typical of concept \( c_d \) should be selected. Typicality is evaluated by rehearsing in the internal process of the system the movement of each possible candidate object, without moving the actual object, and calculating the likelihood of the HMM for \( c_d \) with regard to the trajectory generated in the rehearsal process.

These two principles are integrated in a probabilistic framework.

### Table 1: The concepts taught in the experiments

<table>
<thead>
<tr>
<th>attribute</th>
<th>concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>static</td>
<td>right</td>
</tr>
<tr>
<td></td>
<td>bottom</td>
</tr>
<tr>
<td></td>
<td>yellow</td>
</tr>
<tr>
<td>dynamic</td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>rotate</td>
</tr>
</tbody>
</table>

4. EXPERIMENTS

The camera unit contained three separate CCDs so that three-dimensional information about the scenes could be obtained. The system’s attention was restricted to objects within 90 cm of the camera unit. The robot arm had seven degrees of freedom and the hand had one. A close-talk microphone was used for speech input. The speech observation was represented by using Mel-scale cepstrum coefficients and their delta parameters (twenty-five dimensional). The visual observation was represented by using the such features as position on the table (two-dimensional: horizontal and vertical coordinates), velocity (two-dimensional), color (three-dimensional: \( L^*a^*b^* \) parameters), and size (two-dimensional: width and height). The information of the position in depth coordinate was used in the attention control process, but was not used as the features for the visual observation. A normal-Wishart \( p.d.f. \) was used as the \textit{a priori} \( p.d.f. \) of the mean vector and the covariance matrix of the normal \( p.d.f. \) for each static concept. A male participant using eleven stuffed and clay toys as objects taught language to the system under acoustic conditions typical of an office environment. A rectangular block was put on the right-hand side on the table and was fixed there.

In the first step, fifteen words (nine words for static concepts, and six words for dynamic concepts) were taught (Table 1). These words were taught in two-hundred-twenty-five learning strokes. In each stroke, either a concept about the static characteristics (position, color, size, and their combinations) was taught by uttering a word and directing system’s attention to a stationary object, or else a concept about dynamic characteristics (movement) was taught by uttering a word and moving an object. The lexicon was learned after these strokes were completed, and Fig. 3 shows the values of the mutual information in the cases when the number of clusters was changed in the clustering process of the data set consisting of static visual observations and the speech associated with them. We can see that when the number of clusters was equal to the number of words the person taught for static concepts (9), the mutual information was almost maximized. When the data was divided into nine clusters, the utterances in each of all clusters were same word : the utterances of each of nine words formed one cluster. The learned values of the parameters in the normal \( p.d.f.s \) for static concepts are listed in Table 2. We can see that the variance values were small for the components which are important to represent each concept. The values of the output probability density parameters of each state in HMMs for dynamic con-
Table 2: Values of parameters in the normal p.d.f. for the static concepts.

<table>
<thead>
<tr>
<th>concept</th>
<th>pos. x</th>
<th>pos. y</th>
<th>L*</th>
<th>a*</th>
<th>b*</th>
<th>width</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>µ</td>
<td>σ*</td>
<td>µ</td>
<td>σ*</td>
<td>µ</td>
<td>σ*</td>
<td>µ</td>
</tr>
<tr>
<td>yellow</td>
<td>57</td>
<td>695</td>
<td>32</td>
<td>305</td>
<td>96</td>
<td>10</td>
<td>-17</td>
</tr>
<tr>
<td>red</td>
<td>53</td>
<td>905</td>
<td>27</td>
<td>253</td>
<td>44</td>
<td>10</td>
<td>58</td>
</tr>
<tr>
<td>blue</td>
<td>55</td>
<td>658</td>
<td>30</td>
<td>309</td>
<td>51</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>left</td>
<td>91</td>
<td>100</td>
<td>30</td>
<td>158</td>
<td>67</td>
<td>493</td>
<td>12</td>
</tr>
<tr>
<td>right</td>
<td>21</td>
<td>84</td>
<td>33</td>
<td>184</td>
<td>69</td>
<td>457</td>
<td>5</td>
</tr>
<tr>
<td>high-position</td>
<td>56</td>
<td>825</td>
<td>9</td>
<td>10</td>
<td>66</td>
<td>445</td>
<td>12</td>
</tr>
<tr>
<td>low-position</td>
<td>51</td>
<td>579</td>
<td>53</td>
<td>22</td>
<td>71</td>
<td>382</td>
<td>4</td>
</tr>
<tr>
<td>big</td>
<td>58</td>
<td>728</td>
<td>36</td>
<td>227</td>
<td>61</td>
<td>411</td>
<td>18</td>
</tr>
<tr>
<td>small</td>
<td>53</td>
<td>683</td>
<td>32</td>
<td>335</td>
<td>75</td>
<td>351</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3: Values of the output probability density parameters of each state in the HMM for dynamic concepts.

<table>
<thead>
<tr>
<th>concept</th>
<th>state No</th>
<th>position x</th>
<th>position y</th>
<th>velocity x</th>
<th>velocity y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>µ</td>
<td>σ*</td>
<td>µ</td>
<td>σ*</td>
</tr>
<tr>
<td>move-onto-the-block</td>
<td>1</td>
<td>45</td>
<td>206.2</td>
<td>43</td>
<td>154.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>59</td>
<td>86.6</td>
<td>23</td>
<td>39.0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>77</td>
<td>7.7</td>
<td>35</td>
<td>38.9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>81</td>
<td>0.6</td>
<td>46</td>
<td>7.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>81</td>
<td>0.5</td>
<td>51</td>
<td>0.3</td>
</tr>
<tr>
<td>write-pentagram</td>
<td>1</td>
<td>54</td>
<td>301.3</td>
<td>37</td>
<td>131.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>43</td>
<td>331.7</td>
<td>43</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>48</td>
<td>408.2</td>
<td>33</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>48</td>
<td>327.1</td>
<td>46</td>
<td>49.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>43</td>
<td>256.5</td>
<td>36</td>
<td>115.2</td>
</tr>
</tbody>
</table>

In the next step, grammar was taught by uttering multiple words as a sentence while moving objects. The utterances in the experiments were rather simple. For example, the person put down the red small object which he had lifted, and spoke "lift red small put-down." The following probabilities were learned:

\[ P([S] \rightarrow [\text{object}][\text{action}]), \]
\[ P([S] \rightarrow [\text{action}][\text{object}]), \]
\[ P([S] \rightarrow [\text{object}]), \]
\[ P([S] \rightarrow [\text{action}]). \]

During the teaching of grammar, the system learned the grammar which the person used. The learning was robust against early errors in the learning process the same as our earlier algorithm that used only graphical objects displayed on a screen [14].

The speech understanding function was tested under limited conditions. In the case where red and blue objects were on the table, when the person said "move-onto-the-block," the system selected the red object and put it on the block because \( P((\text{red}, [\text{object}]), (\text{move-onto-the-block}, [\text{action}])) \) was larger than \( P((\text{blue}, [\text{object}]), (\text{move-onto-the-block}, [\text{action}])) \). In addition, in the case where one red object was just on the table and another red object was on the block, when the person said "red move-onto-the-block," the system selected the red object just on the table and put

Figure 3: The change of the values of mutual information between speech and visual information according to the number of clusters.

concepts move-onto-the-block and write-pentagram are listed in Table 3. In the HMM for move-onto-the-block, the final state can be considered to represent the location on the block. The variance values for the position components in the final state were small desirable. In the HMM for write-pentagram, the variance values in the velocity components of each of the states were small. Each state can be considered to represent a line stroke during the drawing of a pentagram. Thus, the concept membership functions were acquired suitably with such a small number of learning samples.
it on the block because the likelihood of the move-onto-the-block HMM with regard to the rehearsed trajectory was larger for the red object just on the table than for the red object already on the block.

5. DISCUSSION

The described lexicon learning method enabled words and concepts to be learned from speech-visual data. The learning problem to be solved here was essentially different from the problem dealt with by the model-selection criteria, such as AIC and MDL. The method would be generally useful for extracting symbolic representation from spatiotemporal continuous information.

In the experiments reported here, the conceptual scene analysis used only two semantic attributes: [action] and [object]. The use of other semantic attributes should also be investigated. Only a small sample of the results obtained were presented in this paper, but a simple learning and inference mechanisms unified in a probabilistic framework made it possible to understand speech and generate actions according to situations. The described algorithm, as a whole, could cope with the problem of open vocabulary and could make the meanings of words to ground to the physical world.

The basic framework would be expandable to realize more natural and flexible language acquisition systems, and the following is a short list of some of the requirements for further improvement of the algorithm:

- Simultaneous learning of lexicon and grammar.
- Treatment of continuous speech including function words.
- Treatment of speech which does not describe the current scene.
- Acquisition of semantic attributes.
- Acquisition of new concepts by using the concept acquired before.
- Attention sharing between a person and a system in a natural way.

6. CONCLUSION

The algorithm based on knowledge representation by graphical statistical model and on a robot-human interface was described. The feasibility of the proposed approach and the requirements for further improvement were shown. Future work will include the expansion of the algorithm and formal evaluation of its performance.

7. ACKNOWLEDGEMENT

I thank Mr. M. Tamura, Mr. A. Miyata, Mr. T. Haoka, and Mr. A. Toyoizumi and Mr. K. Kim for their cooperative work.

8. REFERENCES

[16] Kim, K. and Iwahashi, N.: to be submitted