Learning Place-Names from Spoken Utterances and Localization Results by Mobile Robot

Ryo Taguchi [1], *Yuji Yamada* [1], *Koosuke Hattori* [1], *Taizo Umezaki* [1], *Masahiro Hoguro* [2], *Naoto Iwahashi* [3], *Kotaro Funakoshi* [4], *Mikio Nakano* [4]

[1] Nagoya Institute of Technology, Japan
[2] Chubu University, Japan
[3] National Institute of Information and Communications Technology, Japan
[4] Honda Research Institute Japan Co., Ltd. , Japan

tag@nitech.ac.jp

Abstract

This paper proposes a method for the unsupervised learning of place-names from pairs of a spoken utterance and a localization result, which represents a current location of a mobile robot, without any priori linguistic knowledge other than a phoneme acoustic model. In previous work, we have proposed a lexical learning method based on statistical model selection. This method can learn the words that represent a single object, such as proper nouns, but cannot learn the words that represent classes of objects, such as general nouns. This paper describes improvements of the method for learning both a phoneme sequence of each word and a distribution of objects that the word represents.

Index Terms: Lexical learning, language acquisition, model selection.

1. Introduction

Dialog systems such as communication robots need to have linguistic knowledge. However, the developers cannot describe all knowledge in advance because such systems can be used in situations other than what the developers assume. Therefore, it is desirable that systems automatically learn such knowledge through interactions with users. In particular, household robots have a lot of opportunities to encounter unknown objects. In order to recognize and utter words indicating them, they have to be able to learn correct phoneme sequences and the meanings of the words from user utterances.

It is very difficult to obtain the correct phoneme sequences of unknown words included in user utterances. In previous work [1], phoneme sequences are obtained using pre-defined fixed expressions in which unknown words are inserted, such as "my name is <*name*>", where any name can replace <*name*>. Gorin et al. [2], Alshawi [3] and Roy and Pentland [4] have conducted experiments to extract semantically useful phoneme sequences from natural utterances but they have not yet been able to acquire the correct phoneme sequences with high accuracy. On the other hand, in the field of speech recognition, several methods to extract out-of-vocabulary (OOV) words from utterances by learning and using acoustic and grammatical models of classes of OOV words, such as personal names or place names, have been proposed [5,6,7]. These studies, however, have not been able to improve the accuracy for each phoneme sequence of OOV words by integrating recognition results of multiple utterances.

In previous work [8], we have proposed a lexical learning method based on statistical model selection. This method can acquire phoneme sequences of object names with about 85 percent accuracy. It can learn the only words that represent discrete categories such as object IDs which are given as results of object recognition. Namely, objects must be classified before lexical learning.

This paper proposes a method to be able to classify objects in parallel with lexical learning. In the experiments, a mobile robot learns ten place-names from pairs of a spoken utterance and a localization result, which represents the current location of the robot, without any priori linguistic knowledge other than a phoneme acoustic model.

2. Lexical learning task

This paper deals with the task where a robot learns names of objects or places. When a user teaches a name of an object to the robot, the user speaks about the object while showing the object to the robot. On the other hand, when the user teaches a place-name, the user brings the robot to each place and speaks about the place. The robot gains a feature vector from the object or the place as its sensor output. Therefore, in the following description, we explain the method considering places as objects. The user uses natural expressions for the instructions. User utterances may include words other than names of objects (or places). For example, the user might say "this is James." In this paper, names of objects are called *keywords*, and words (or phrases) other than keywords are called *non-keyword expressions*. We assume that the keywords and the non-keyword expressions are independent of each other. Therefore, the same non-keyword expressions can be used in instruction utterances for different keywords. The robot has never been given linguistic knowledge other than an acoustic model of Japanese phonemes. By using the acoustic model, it can recognize user utterances as Japanese phoneme sequences, but cannot extract keywords from them. The robot must learn the correct phoneme sequences and the meanings of keywords from a set of pairs of an utterance and an object. After learning, we estimate the learning result by investigating whether it can output the correct phoneme sequence corresponding to each object.

3. Approach

The main problems of the above task are word boundary detection from the utterances, and phoneme sequence estimation of the words. Since the phoneme sequences obtained by recognizing utterances may contain errors, it is

difficult to correctly identify the word boundaries. For example, Roy and Pentalnd [4] has extracted keywords by using similarities of both acoustic features and meanings, but 70% of the extracted words contained insertion or deletion errors at either or both ends of the words. Moreover, this method did not have a mechanism to select the most correct phoneme sequence for each word from a lot of word candidates obtained through learning.

In order to solve the problems, in our method, a statistical model of the joint probability of an utterance and an object is learned based on the minimum description length principle. This model consists of a word list, in which each word is represented by a phoneme sequence, and three statistical models: the phoneme acoustic model, a word-bigram model, and a word meaning model. We call this model the utterance-object joint probability model. The phoneme acoustic model has been learned beforehand because it requires much more speech data. By alternating between the learning of the other two statistical models and the optimization of the word list, acoustically, grammatically and semantically appropriate phoneme sequences are acquired as words.

4. Utterance-object joint probability model

Joint probability $P(\mathbf{a}, \mathbf{o})$ of spoken utterance **a** and an object **o** is expressed in Eq. 1.

$$\begin{aligned} P(\mathbf{a},\mathbf{o}) &= \sum_{s} P(\mathbf{a},\mathbf{o},s) \\ &= \sum_{s} \{P(\mathbf{a} \mid s)P(s)P(\mathbf{o} \mid s)\} \end{aligned} \quad (1)$$

Here, **a** is a feature vector extracted from an utterance. Object **o** is a m-dimensional vector representing an object (or a place). s is a word sequence which consists of n words $(w_1,\ldots,w_n)$, start point w_0 and end point w_{n+1}. Some of the words included in s are keywords and the others are non-keyword expressions.

$P(\mathbf{a} \mid s)$ represents an acoustic score which is the likelihood of the word sequence s, which is calculated using the phoneme acoustic model as usual speech recognition systems do. $P(s)$ represents a grammatical score which is calculated from the word-bigram language model. Note that we use a class bigram model in which all keywords are treated as words in the keyword class. $P(\mathbf{o} \mid s)$ represents a semantic score which is described in section 4.1.

Equation 1 needs a huge amount of calculation because there are a huge number of word sequences. Therefore, we approximate the summation by maximization. Moreover, in order to adjust differences of the accuracies of the three statistical models, we multiply the acoustic score by the weighting parameter α (0.0001 in this work). Therefore the logarithm of utterance-object joint probability is defined as follows:

$$\begin{aligned} &\log P(\mathbf{a},\mathbf{o}) \\ &\quad \approx \max_{s} \{\alpha \log P(\mathbf{a} \mid s) + \log P(s) + \log P(\mathbf{o} \mid s)\} \end{aligned} \quad (2)$$

4.1. Calculation of Semantic Score

Semantic score $P(\mathbf{o} \mid s)$ is defined as follows:

$$P(\mathbf{o} \mid s) = \sum_{i=1}^{n} \gamma(s,i) P(\mathbf{o} \mid w_i) \quad (3)$$

$$\gamma(s,i) = \frac{\mathrm{N}(w_i)}{\mathrm{N}(s)} \quad (4)$$

where $\gamma(s,i)$ is a weighting function, $N(w_i)$ is the number of phonemes of word w_i, $N(s)$ is the total amount of phonemes of keywords included in a word sequence s. $\gamma(s,i)$ is assigned zero when word w_i is not a keyword. $P(\mathbf{o}|w)$ is a word meaning. This means the meaning of an utterance is inferred from keywords in the utterance.

In order to determine whether or not a word is a keyword, the difference between the entropy of object **o** and the conditional entropy of object **o** given w is calculated as follows:

$$\begin{aligned} I(w) = &- \int P(\mathbf{o}) \log P(\mathbf{o})\, d\mathbf{o} \\ &+ \int P(\mathbf{o} \mid w) \log P(\mathbf{o} \mid w)\, d\mathbf{o} \end{aligned} \quad (4)$$

If the difference $I(w)$ is higher than a certain threshold T, the word w is considered as a keyword. T was manually determined on the basis of preliminary experimental results.

4.2. Word Meaning Model

Previous work [8] has used discrete probability distribution as the model of $P(\mathbf{o}|w)$. In this paper, we use multidimensional Gaussian distribution as the model in order to perform classification of objects.

$$P(\mathbf{o} \mid w) = \frac{1}{(\sqrt{2\pi})^m \sqrt{|\mathbf{S}|}} \exp\left(-\frac{1}{2} (\mathbf{o}-\boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{o}-\boldsymbol{\mu}) \right) \quad (5)$$

where m is the dimension number of **o**, **S** is the variance-covariance matrix and **μ** is the mean vector of **o** given w. In the experiments described in section 6, we use 2-dimensional Gaussian distribution for learning several places. However, this method may have applicability to other feature vectors such as the shape of an object and its color.

4.3. Keyword Output

When the robot receives an object **o**, it outputs the keyword $w_{\mathbf{o}}$ that is the best to represent the object **o** based on Eq. (6).

$$\begin{aligned} w_{\mathbf{o}} &= \arg\max_{w \in \Omega} P(w \mid \mathbf{o}) \\ &= \arg\max_{w \in \Omega} P(w, \mathbf{o}) \\ &= \arg\max_{w \in \Omega} \left\{ \log P(w) + \log P(\mathbf{o} \mid w) \right\} \end{aligned} \quad (6)$$

where Ω is the set of acquired keywords.

5. Lexical learning method

Figure 1 gives an overview of our method for lexical learning. It consists of three steps, namely building the initial word list, the learning of the word-bigram language model and the word meaning model, and the optimization of the word list based on a model selection technique.

5.1. Step1: Building of the initial word list

At first, all user utterances are recognized as phoneme sequences by using the phoneme acoustic model. Next, a word list is built by extracting subsequences included in the phoneme sequences. The entropies of phonemes before or after each subsequence are calculated. If the entropies of a subsequence are not zero and the frequency of the subsequence is more than two, the subsequence is registered on the word list as a word candidate.

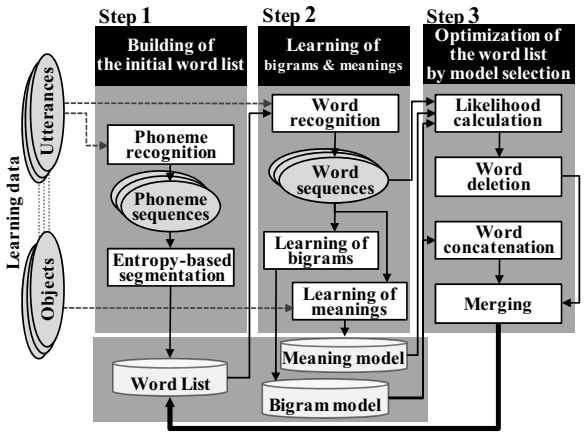

Figure 1: *Overview of the lexical learning method.*

5.2. Step2: Learning of the word-bigram language model and the word meaning model

The utterances are re-recognized as word sequences using both the phoneme acoustic model and the word list. Note that N-best hypotheses are output as a recognition result for each utterance in this system. The word-bigram language model and the word meaning model are learned from the word sequences included in the N-best hypotheses.

5.3. Step3: Optimization of the word list by model selection

The number of words in the word list is optimized based on the minimum description length principle [9,10] which is the basis for model selection. In this paper, we define the description length *DL( )* as follows:

$$DL(\theta) = -L(\theta, \mathbf{D}) + \frac{f(\theta)}{2}\log M \tag{4}$$

$$L(\theta, \mathbf{D}) = \sum_{i=1}^{M} \log P(\mathbf{a}_i, \mathbf{o}_i; \theta) \tag{5}$$

$$f(\theta) = K + (K^2 + 2K) + (K(m + m(m+1)/2)) \tag{6}$$

where $L(\theta, \mathbf{D})$ is a log likelihood of θ, $\mathbf{D} = \{d_i \mid 1 \le i \le M\}$ is the set of learning data $d_i = (\mathbf{a}_i, \mathbf{o}_i)$. $f(\theta)$ is the number of parameters of the word-bigram language model and the word meaning model, and represents the degrees of freedom of θ. *K* is the number of words and *m* is the dimension number of **o**.

The optimization of the word list requires calculating the log likelihoods in all combinations of possible word candidates. However, it is computationally expensive and not practical. Therefore, using the N-best hypotheses obtained in Step 2, we approximately calculate the difference of the description lengths of two models, one that includes word *w* and the other that does not. This is done by computing the likelihood of the hypothesis that is the highest among ones that do not include word *w*. The description length of the model θ_1 obtained by subtracting word *w* from the original model θ_0, $DL(\theta_1)$, is calculated by subtracting the difference from $DL(\theta_0)$. If $DL(\theta_1)$ is lower than $DL(\theta_0)$, word *w* is removed from the original model θ_0. This word deletion is iterated in order of decreasing difference of DLs.

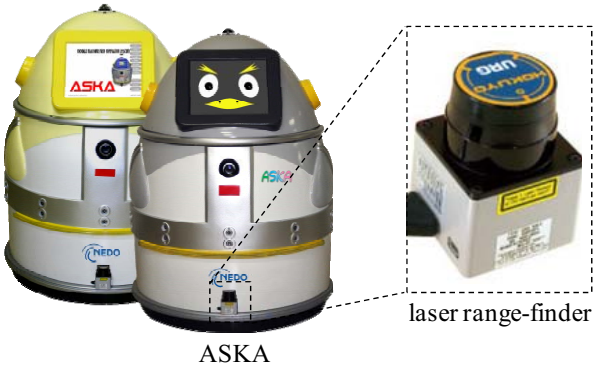

Figure 2: *Mobile robot ASKA.*

In addition, if the bigram probability of two words is higher than a certain threshold (0.3 in this work), a new word candidate is generated by concatenating them into one word. This leads to recovering the erroneous dividing of words in Step 1.

A new word list is built by merging both results. After that, the process returns to Step 2 and re-learns the word bigram-language model and the word meaning model by using the new list. Through the iteration of these processes, acoustically, grammatically, and semantically useful words are acquired.

However, in the early phase of iteration, important words may be deleted too much. In order to solve this problem, in the early phase of iteration, this method just executes word concatenation but doesn't delete words. In the following experiment, word deletion starts after the second iteration.

6. Experimental results and discussions

6.1. Experimental Conditions

We incorporated the above learning method into the mobile robot ASKA which had been developed by our laboratory [11]. Figure 2 shows ASKA. It can construct a map and localize itself in the map. We used FastSLAM algorithm [12] for map construction. Moreover, ASKA estimates its own location by using its laser range-finder. The localization result is represented as a two dimensional vector (*x*, *y*) which represents a position coordinate in the map. In this experiment, the vector (*x*, *y*) is considered as object **o**.

First, we moved ASKA with an infrared remote control and made it construct a map of the second floor of the building where our laboratory is. Next, we moved ASKA to several points on the floor and taught it the name of each place by voice. We taught ten place-names. Nine non-keyword expressions are used such as "kokowa <*keyword*> desu" and "konobasyowa <*keyword*>", where each keyword can replace <*keyword*>. The ninety utterances which consist of all combinations of keywords and non-keyword expressions were recorded. We gave ASKA each of the utterances at a different point on the floor. ASKA learned place-names from ninety pairs of the utterances and localization results.

6.2. Experimental Results

The results of the experiment, in which the iterations of Step 2 and Step 3 were carried out ten times, are shown in Figure 3 and 4. Figure 3 shows the constructed map, the correct phoneme sequences of taught keywords, which are shown in round brackets, and keywords output by ASKA after the 10th iteration. The phoneme accuracy for the keywords was 80%.

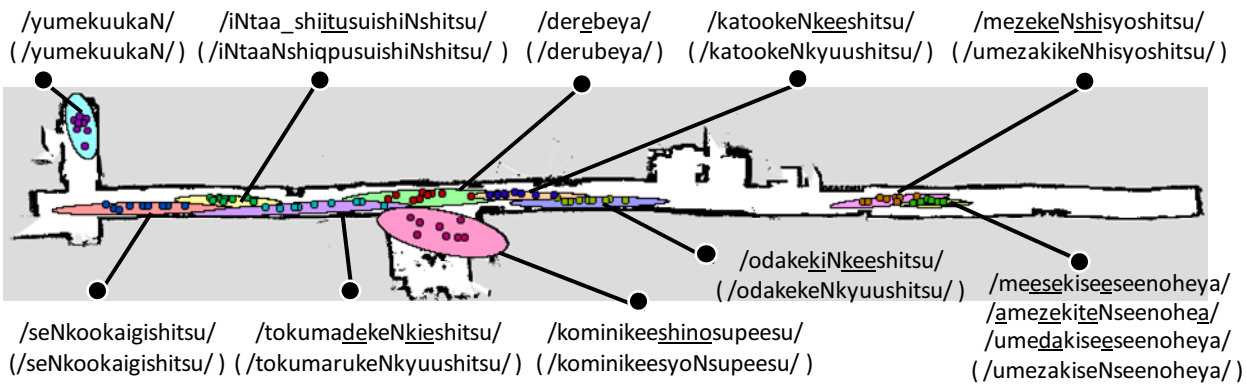

Figure 3: *Output keywords and P*(*w*, **o**) *of each keyword after 10th iteration.*

This performance is approximately the same as the previous work [8]. In the figure, small circles represent the position coordinates that were used for learning, and each big ellipse represents probability *P*(*w,* **o**) of each keyword. These results show that the area represented by each place-name was learned. Namely, places or objects were classified suitably.

Figure 4 shows the number of acquired words and keywords included in the word list. The horizontal axis represents the number of learning iterations. We can see that the iterations made the number of words decrease. Finally, 12 keywords were obtained. This number is close to ten, which is the real number of keywords in the training utterance set.

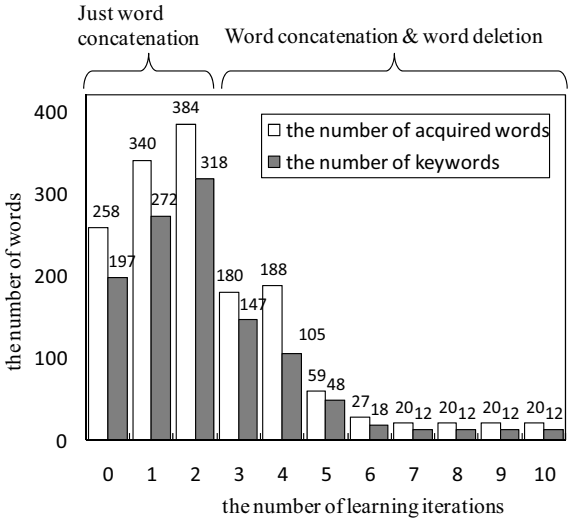

Figure 4: *Effects of the optimization of the word list.*

7. Conclusions

This paper proposed a method for the unsupervised learning of place-names from pairs of a spoken utterance and a localization result without any priori linguistic knowledge other than a phoneme acoustic model. Experimental results show that, without a priori word knowledge, it can acquire phoneme sequences of place-names with about 80 percent accuracy. Moreover, the area represented by each place-name was suitably learned. These results show that this method can classify places or objects in parallel with lexical learning. We expect the basic principle presented in this study will provide us with a clue to resolve the general language acquisition problem in which morphemes of speech language are extracted by using only non-linguistic semantic information related to each utterance.

8. Acknowledgements

This research was supported by Grant-in-Aid for Scientific Research (B) (Subject No. 2130083).

9. References

[1] Holzapfel, H., Neubig, D. and Waibel, A., "A Dialogue Approach to Learning Object Descriptions and Semantic Categories", Robotics and Autonomous Systems, Vol. 56, Issue 11: 1004-1013, 2008.

[2] Gorin, A. L., Petrovska-Delacretaz D., Wright, J. H. and Riccardi, G., "Learning spoken language without transcription", Proc. ASRU Workshop, 1999.

[3] Alshawi, H., "Effective utterance classification with unsupervised phonotactic models", Proc. NAACL 2003.

[4] Roy, D., and Pentland, A., "Learning words from sights and sounds: A computational model", Cognitive Science, 26, 113-146, 2002.

[5] Asadi, A., Schwartz, R. and Makhoul, J., "Automatic Modeling for Adding New Words to a Large Vocabulary Continuous Speech Recognition System", Proc. ICASSP91:305-308, 1991.

[6] Schaaf, T., "Detection of OOV Words Using Generalized Word Models And A Semantic Class Language Model", Proc. Eurospeech 2001.

[7] Bazzi, I., and Glass, J., "A multi-class approach for modelling out-of-vocabulary words", Proc. ICSLP02, 1613-1616, 2002.

[8] Taguchi, R., Iwahashi, N., Nose, T., Funakoshi, K., and Nakano, M., "Learning Lexicons from Spoken Utterances Based on Statistical Model Selection," Proc. of INTERSPEECH2009, 2731-2734, 2009.

[9] Rissanen, J., "A universal prior for integers and estimation by minimum description length", The Annals of Stat., Vol.11, No.2, 416-431, 1983.

[10] Rissanen, J. "Stochastic complexity and modeling", The Annals of Stat., Vol.14, No.3, 1080-1100, 1986.

[11] Hoguro, M., Tsuzuki, H., Ikeyama, J., Umezaki, T., "The Development of the Tracking Robot"ASKA"and Its Application to Security System," Proc. of ISR 2005.

[12] Hahnel, D., Burgard, W., Fox, D., Thrun, S., "An Efficient FastSLAM Algorithm for Generating Maps of Large-Scale Cyclic Environments from Raw Laser Range Measurements," Proc. of the IEEE/RSJ Int. Conf. on IROS, 2003.