Robotics and Autonomous Systems 99 (2018) 166–180

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Unsupervised spatial lexical acquisition by updating a language model with place clues☆

Akira Taniguchi [a,*], Tadahiro Taniguchi [a], Tetsunari Inamura [b]

[a] Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
[b] National Institute of Informatics/The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

HIGHLIGHTS

- We improve the accuracy of lexical acquisition by updating a language model with place clues.
- The robot can learn spatial concepts with high accuracy as unsupervised place categorization.
- The mutual information is used to select words related to a place effectively.

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form 22 September 2017
Accepted 19 October 2017
Available online 2 November 2017

Keywords:
Ambiguous speech recognition
Bayesian nonparametrics
Lexical acquisition
Place categorization
Spatial concept acquisition
Unsupervised word segmentation

ABSTRACT

This paper describes how to achieve highly accurate unsupervised spatial lexical acquisition from speech-recognition results including phoneme recognition errors. In most research into lexical acquisition, the robot has no pre-existing lexical knowledge. The robot acquires sequences of some phonemes as words from continuous speech signals. In a previous study, we proposed a nonparametric Bayesian spatial concept acquisition method (SpCoA) that integrates the robot's position and words obtained by unsupervised word segmentation from uncertain syllable recognition results. However, SpCoA has a very critical problem to be solved in lexical acquisition; the boundaries of word segmentation are incorrect in many cases because of many phoneme recognition errors. Therefore, we propose an unsupervised machine learning method (SpCoA++) for the robust lexical acquisition of novel words relating to places visited by the robot. The proposed SpCoA++ method performs an iterative estimation of learning spatial concepts and updating a language model using place information. SpCoA++ can select a candidate including many words that better represent places from multiple word-segmentation results by maximizing the mutual information between segmented words and spatial concepts. The experimental results demonstrate a significant improvement of the phoneme accuracy rate of learned words relating to place in the proposed method by word-segmentation results based on place information, in comparison to the conventional methods. We indicate that the proposed method enables the robot to acquire words from speech signals more accurately, and improves the estimation accuracy of the spatial concepts.

1. Introduction

Autonomous robots operating in a human living environment need to understand the spatial lexical knowledge of their ambient environment in order to facilitate interactions with people. For example, a robot might be required to recognize the name of its current position and those of other areas on its environmental map, such as *"kitchen"*, *"entrance-way"*, and certain proper nouns specific to various places. Therefore, we consider it to be important for robots to be able to learn the novel and various words that people associate with particular places in their environments, and the spatial areas corresponding to those names, i.e., robots must be able to acquire novel and various words relating to places. To do this, a robot could use speech signals recognized from microphones and the sensory-motor information obtained from odometry and laser sensors in the ambient environment.

Lexical acquisition means that a robot with no pre-existing lexicon learns phoneme sequences from the continuous speech signals of a person. In this case, the robot must be able to manage a considerable degree of uncertainty in the speech recognition, i.e., phoneme recognition errors. When the robot learns novel

☆ This research was partially supported by JST CREST, and a Grant-in-Aid for Scientific Research on Innovative Areas (16H06569) funded by the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

* Corresponding author.
E-mail addresses: a.taniguchi@em.ci.ritsumei.ac.jp (A. Taniguchi), taniguchi@em.ci.ritsumei.ac.jp (T. Taniguchi), inamura@nii.ac.jp (T. Inamura).

https://doi.org/10.1016/j.robot.2017.10.013
0921-8890/

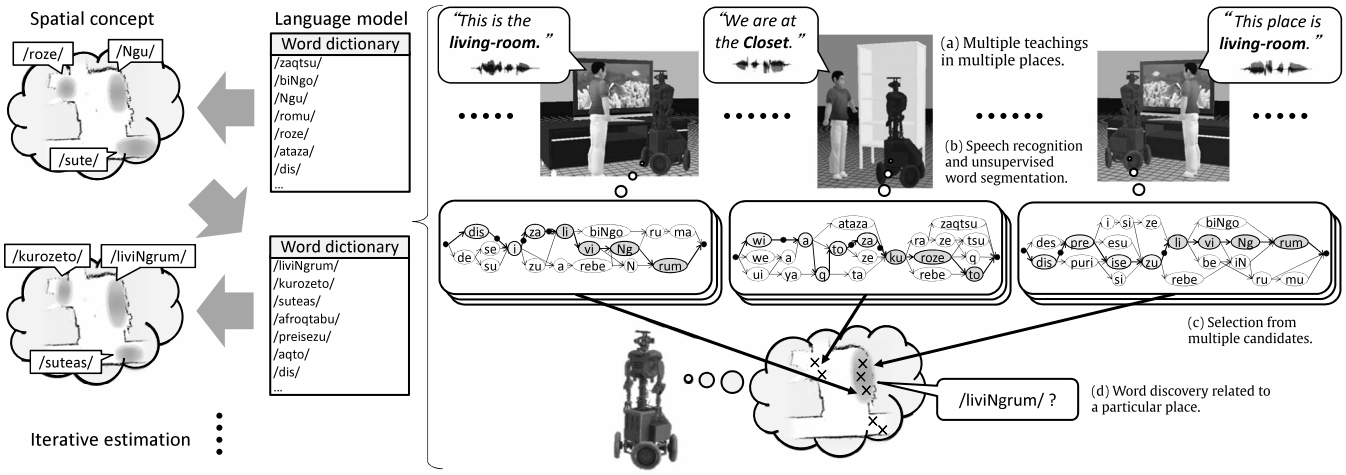

Fig. 1. Schematic representation of spatial lexical acquisition used in this study. Our approach involves the updating of both spatial concepts and a language model for highly accurate word discovery. (a) The user teaches the robot different sentences in different places. (b) The robot recognizes the user's speech signals and segments them into words. The thickest line in each weighted finite-state transducer (WFST) represents the selected path, and a black dot in the middle of an arrow represents a word boundary. For example, in the WFST on the right, the words that were obtained are */dis/*, */preisezu/*, and */liviNgrum/*. (c) The robot selects words from multiple candidates of segmented words by using place information. (d) The robot learns words related to a particular place.

words from speech signals, it is difficult to determine segmentation boundaries and the identity of words by using speech recognition, which can lead to various errors. Without pre-existing lexical knowledge, the robot makes speech-recognition mistakes more frequently than if it had the required lexical knowledge [1]. Let us consider a problem related to lexical acquisition from uttered sentences. For example, the robot obtains a speech-recognition result such as */heaiznyuyoak/* (an incorrect phoneme recognition of *"Here is New York"*.). The robot must segment the sentence into the true boundaries as individual words, e.g., */hea/*, */iz/*, and */nyuyoak/*. However, in many cases, this speech-recognition result is segmented into incorrect boundaries through either under- or over-segmentation [2], e.g., */he/*, */aiz/*, */nyuyo/*, and */ak/*. Furthermore, it is necessary for the robot to recognize words referring to the same thing from among these numerous segmented results that contain errors.

This study addresses the above lexical-acquisition problems by complementing ambiguous speech-recognition and word-segmentation results with place information. We assume that the robot has not acquired any vocabulary in advance, and can recognize only phonemes or syllables. We represent the spatial area of the environment in terms of a *position distribution*. Furthermore, we define a *spatial concept* as a place category that includes place names and the position distributions corresponding to those names.

Taniguchi et al. [3] proposed nonparametric Bayesian spatial concept acquisition method (SpCoA) based on an unsupervised word-segmentation method known as latticelm[1] [4]. This method enables word segmentation with consideration of phoneme errors in speech recognition more efficiently than does the nested Pitman–Yor language model (NPYLM) [5]. However, in many cases, the original word representing the name of the place is finely segmented into several words, i.e., over-segmentation. We consider this problem to be caused by a word-segmentation method that does not use place information, i.e., the words are segmented from syllable sequences only. In this paper, as a solution to this problem, we propose the SpCoA++ method that iteratively constructs spatial concepts and a language model, and that performs word segmentation and spatial concept acquisition more accurately. In addition, we propose a method that enables a word related to a particular place to be selected using the mutual information.

A schematic diagram depicting our study is shown in Fig. 1. The left parts of Fig. 1 represent the iterative estimation of spatial concepts and a language model. In Fig. 1(a), when the robot arrives at a place that is a designated learning target, the user speaks a sentence (including the name of the place) to the robot, which moves within the environment while performing self-localization. In Fig. 1(b), the robot performs speech recognition and unsupervised word segmentation from the human speech signals. In our approach, in order to cope with the uncertainty of speech recognition, we use speech recognition based on a weighted finite-state transducer (WFST), which is a word graph representing the speech-recognition results. In addition, we use the latticelm [4], which can segment the speech-recognition results in the WFST format. In Fig. 1(c), from the multiple candidates of word boundaries and paths in a WFST, the robot selects the one that includes words that best represent a particular place. In Fig. 1(d), the robot learns a word that is frequently obtained in only a specific place as the name of the place, e.g., */liviNgrum/*. The robot then uses the selected words to update a language model. We consider that our approach can greatly improve the performance of unsupervised word segmentation from speech-recognition results that contain phoneme-recognition errors in the lexical acquisition related to places.

The main contributions of this paper are as follows.

- We improve the accuracy of spatial lexical acquisition by updating a language model with place clues. In comparative experiments, the proposed method performs better than the conventional methods.
- We show that it is possible to learn spatial concepts with increased accuracy as unsupervised place categorization by obtaining highly accurate word segmentation.
- We show that it is possible to select word-segmentation results and words related to a particular place by using the mutual information. In addition, we show that the selection of word-segmentation candidates by using this mutual information is theoretically valid.

The remainder of this paper is organized as follows. In Section 2, we discuss previous studies on lexical acquisition and semantic mapping that are relevant to our study. In Sections 3 and 4, we

[1] latticelm is an unsupervised word-segmentation tool that [4] is implemented and is treated as the name of the method in this study.