Evaluation of a System for Noun Concepts Acquisition from Utterances about Images (SINCA) Using Daily Conversation Data

Yuzu UCHIDA
Graduate School of
Information Science and Technology
Hokkaido University
Sapporo, 060-0814, Japan
yuzu@media.eng.hokudai.ac.jp

Kenji ARAKI
Graduate School of
Information Science and Technology
Hokkaido University
Sapporo, 060-0814, Japan
araki@media.eng.hokudai.ac.jp

Abstract

For a robot working in an open environment, a task-oriented language capability will not be sufficient. In order to adapt to the environment, such a robot will have to learn language dynamically. We developed a System for Noun Concepts Acquisition from utterances about Images, SINCA in short. It is a language acquisition system without knowledge of grammar and vocabulary, which learns noun concepts from user utterances. We recorded a video of a child's daily life to collect dialogue data that was spoken to and around him. The child is a member of a family consisting of the parents and his sister. We evaluated the performance of SINCA using the collected data. In this paper, we describe the algorithms of SINCA and an evaluation experiment. We work on Japanese language acquisition, however our method can easily be adapted to other languages.

1 Introduction

There are several other studies about language acquisition systems. Rogers et al. (1997) proposed "Babbette", which learns language rules from provided examples. Levinson et al. (2005) describe their research with a robot which acquires language from interaction with the real world. Kobayashi et al. (2002) proposed a model for child vocabulary acquisition based on an inductive logic programming framework. Thompson (1995) presented a lexical acquisition system that learns a mapping of words to their semantic representation from training examples consisting of sentences paired with their semantic representations.

As mentioned above, researchers are interested in making a robot learn language. Most studies seem to be lacking in the ability to adapt to the real world. In addition, they should be more independent from language rules. We believe that it is necessary to simulate human language ability in order to create a complete natural language understanding system.

As the first step in our research, we developed a System for Noun Concepts Acquisition from utterances about Images, called SINCA in short (which means "evolution" in Japanese) (Uchida et al., 2007). It is a language acquisition system without knowledge of grammar and vocabulary, which learns noun concepts from a user's input. SINCA uses images as a meaning representation in order to eliminate ambiguity of language. SINCA can only acquire concrete nouns.

Currently, SINCA is for Japanese only. The language acquisition method of this system is very general and it is independent of language rules. SINCA is expected to work successfully using any language.

In this paper, we describe the algorithms of SINCA and an experiment to test what kind of input would be appropriate for our system. We would emphasize that we prepared a large video data of daily life of a family with young children.

2 The Algorithms of SINCA

Figure 1 shows the SINCA user interface. The situation shown in Fig.1 is that the affection of SINCA is directed to an eraser by the user, and after the recognition process, SINCA asks *"KESHIGOMU?*

Proceedings of NAACL HLT 2009: Short Papers, pages 65–68,
Boulder, Colorado, June 2009.

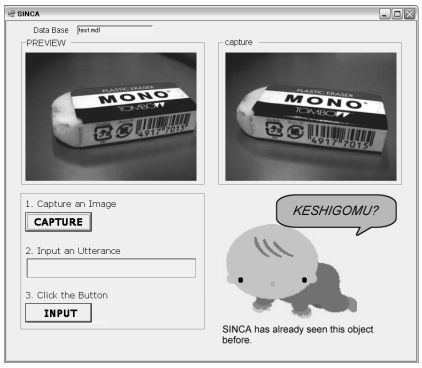

Figure 1: The SINCA Interface recognizing an eraser

*Kore-ha **KAPPU**-tte iu-n-da-yo.*
(This is a thing called a cup.)
***KAPPU**-ni gyūnyū ireyōka.*
(Let's pour some milk into the cup.)

Strings indicated by boldface are labels.

Figure 2: Examples of input data

(Eraser?)."

We describe SINCA's process in detail in the following subsections.

2.1 Input

A user input consists of an image paired with a spoken utterance.

First, a user chooses an object *O* which he or she likes and captures an image of it with a web camera with 300,000 pixels effective sensor resolution. The user has to try to capture the whole object *O* in the image.

Next, a user imagines an utterance that an infant might be exposed to when listening to caregivers while gazing at the object *O* in the environment. The user enters the utterance on the keyboard as a linguistic input. The linguistic input is written in *Hiragana*, which are Japanese phonetic characters, to avoid the linguistic input containing some direct meanings as in the case of Chinese *Kanji* ideograms. This is also intended to standardize the transcription. SINCA does not carry out morphological analysis of the linguistic input, because we believe that infant capability for word segmentation is not perfect (Jusczyk et al., 1999).

Figure 2 shows some example inputs. [1]

2.2 Image Processing

The ERSP 3.1 Software Development Kit [2] provides cutting edge technologies for vision, navigation, and system development. ERSP Vision included in the ERSP enables a robot or device to recognize 2D and 3D objects in real world settings where lighting and placement are not controlled. We use the ERSP vision for image processing. ERSP Vision informs the system whether the object in the present input image appears in the previously input images or not.

2.3 Common Parts

When a user inputs an image of an object *O* and an utterance, the system extracts all sections of the string matching section of previously input utterances accompanied by the image of the same object *O*. We call these strings common parts. After this process, the system deals with them as candidates for a label for the object *O*.

The system provides every common part with a "basic score". The basic score is based on frequency of appearance and the number of characters, and indicates how appropriate as a label the common part is. The higher the score, the more appropriate the common part is. The basic score is defined as follows:

$$SCORE = \alpha \times \frac{F}{PN} \times \sqrt{L} \qquad (1)$$

where, α is a coefficient which reduces the basic score if the common part has appeared with other objects than *O*, F is frequency of appearance of the common part with the images of *O*, PN is the number of use inputs with images of *O*, and L is the number of characters of the common part.

[1] The Japanese words are written in italics in all following figures.

[2] Evolution Robotics, Inc.:ERSP 3.1 Robotic Development Platform OEM Software by Evolution Robotics

2.4 Output

If the system finds a common part whose basic score exceeds a threshold, it outputs it as text. The reason for doing this is the assumption that there is a high possibility that such common parts are appropriate as labels.

A user evaluates an output by choosing one of the following keywords:

- Good : It is appropriate as a label.
- Almost : It makes some sense but is not proper for the label.
- Bad : It makes no sense.

Infants cannot understand these keywords completely, but they can get a sense of some meanings from the tone of an adult's voice or facial expressions. In our research, we use the keywords as a substitute for such information. The system recalculates the basic score based on the keyword chosen by the user. Specifically, the system multiplies the basic score by the coefficient β dependent on the keyword.

2.5 Acquisition of the Noun Concepts

After repeating these processes, if there is a common part whose score is more than 30.0 and which has been rated as "Good", the system acquires the common part as the label for *O*.

2.6 Label Acquisition Rules

Humans can use their newfound knowledge to learn their native language effectively. This system imitates humans' way with "label acquisition rules".

A label acquisition rule is like a template, which enables recursive learning for acquisition of noun concepts. The system generates label acquisition rules after acquisition of a label. When the system acquires a string S as a label for an object, the system picks up the previous linguistic inputs with the images of the object which contain the string S. Then, the system replaces the string S in the linguistic inputs with a variable "γ". These abstracted sentences are called label acquisition rules. An example of the label acquisition rules is shown in Fig.3.

If the rules match other parts of previoiusly input strings, the parts corresponding to the "γ" variable are extracted. The scores of these extracted strings are then increased.

Acquired Label	: ***WAN-CHAN*** (a doggy)
Previous Input	: *Acchi-ni* ***WAN-CHAN****-ga iru-yo.* (There is a doggy over there.)
Label Acquisition Rule	: *Acchi-ni* γ***1****-ga iru-yo.* (There is γ1 over there.)

Strings indicated by boldface are labels.

Figure 3: An example of a label acquisition rule

3 Evaluation Experiment

We carried out an experiment to test what kinds of input would be appropriate for SINCA. This section describes the experiment.

3.1 Experimental Procedure

Two types of linguistic input data were collected in two different ways: a questionnaire and a video recording. We had SINCA acquire labels for 10 images using the linguistic input data. The following are the details about the data collection methods.

3.1.1 Questionnaire

10 images were printed on the questionnaire, and it asked "What would you say to a young child if he or she pays attention to these objects?". The respondents are allowed to answer with whatever they come up with. 31 people responded to this questionnaire, and 13 of them have children of their own. We collected 324 sentences, and the average mora length of them was 11.0.

3.1.2 Video recording

We recorded a video of a child's daily life to collect dialogue data that was spoken to and around him. The child is a member of a family consisting of his parents and his sister.

The recordings are intended to collect daily conversation, therefore we did not set any tasks. The total recording period comprised 125 days and we recorded about 82 hours of video data. The first author watched about 26 hours of the video data, and wrote parents' dictation in *Hiragana*. We selected 353 sentences for linguistic input data that were spoken when joint attention interactions between a parent and a child were recognized. On average, their mora length was 9.8.

3.2 Experimental Result

We input sentences from the collected inputs one at a time until SINCA acquired a noun concept for an image. SINCA was able to acquire labels for 10 images, with each type of linguistic input. When we used the questionnaire data, SINCA needed on average 6.2 inputs to acquire one label, and SINCA acquired 52 rules through the experiment. They cover 83.8% of the total number of inputs. When we used the video data, SINCA needed on average 5.3 inputs to acquire one label, and SINCA acquired 44 rules through the experiment. They cover 83.0% of the total number of inputs.

3.3 Considerations

The experimental results indicate that using video data makes the acquisition of labels more efficient. There are 3 factors that contribute to this.

The first factor is the number of one-word sentences. There are 66 one-word sentences in the video data (18.6% of the total). Therefore, the length of the sentences from the video data tends to be short.

The second factor is the lack of particles. The respondents of the questionnaire hardly ever omit particles. By contrast, of the 53 sentences which were input, 23 sentences lack particles (42.6% of the total) in video data. Spoken language is more likely to have omitted particles compared with written language.

The third factor is the variety of words. We randomly selected 100 sentences from both sets of linguistic input data and checked the words adjacent to a label. Table 1 shows the number of different words that occur adjacent to a label. Because the respondents of the questionnaire all try to explain something in an image, they use similar expressions.

When SINCA uses the video data, it can extract labels more easily than using the questionnaire data because of the factors listed above. This means that SINCA is well suited for spoken language. If we assume one application of SINCA is for communication robots, this result is promising.

Table 1: Variety of words

	Previous(W_A)	following(W_B)
Video	19	42
Questionnaire	15	22

$Sentence : W_1\ W_2\ ...\ W_A$ label $W_B\ ...$.

4 Conclusions and Future Work

In this paper, we described the algorithms of SINCA. SINCA can acquire labels for images without ready-made linguistic resources, lexical information, or syntactic rules. Additionally, it targets images of real world objects.

We collected linguistic input data in two ways. One method is videos of a family's daily life. The other method is a questionnaire. We had SINCA acquire noun concepts using both video and questionnaire data. As a result, we have showed that spoken language is well suited to SINCA's algorithm for acquiring noun concepts.

In the next step, we will focus on acquisition of adjectives.

References

Jusczyk, P. W. Houston, D. M. and Newsome, M. 1999. *The beginnings of word segmentation in english-learning infants*. Cognitive Psychology. **39**. pp.159–207.

Kobayashi, I. Furukawa, K. Ozaki, T. and Imai, M. 2002. *A Computational Model for Children's Language Acquisition Using Inductive Logic Programming*. Progress in Discovery Science. **2281** pp.140–155.

Levinson S. E. Squire, K. Lin, R. S. and McClain, M. 2005. *Automatic language acquisition by an autonomous robot*. AAAI Spring Symposium on Developmental Robotics.

Rogers, P. A. P. and Lefley, M. 1997. *The baby project*. Machine Conversations. ed. Wilks, Y. Kluwer Academic Publishers.

Thompson, C, A. 1997. *Acquisition of a Lexicon from Semantic Representations of Sentences*. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. pp.335–337.

Uchida, Y. and Araki, K. 2007. *A System for Acquisition of Noun Concepts from Utterances for Images Using the Label Acquisition Rules*. Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI). pp.798–802.