

Modeling Situated Conversations for a Child-Care Robot Using Wearable Devices

Kyoung-Woon On, Eun-Sol Kim, and Byoung-Tak Zhang

School of Computer Science and Engineering & Cognitive Robotics and Artificial Intelligence
Seoul National University
{kwon, eskim, btzhang}@bi.snu.ac.kr

Abstract

How can robots fluently communicate with humans and have context-preserving conversation? It is the most momentous and crucial problem in robotics research, especially for service robots such as child-care robots. Here, we aim to develop a situated conversation system for child-care robots. The conversation system considers the current context between robots and children as well as the situation the child is in. The system consists of two parts. The first part tries to understand the context. This part uses the embedded sensors of robots to understand the context and wearable sensors of the child for getting information of the situation the child is in. The second part is to generate the situated conversation. In terms of the model, we designed a hierarchical Bayesian Network for the first part and a Hypernetwork model is used for the second part. We illustrate the application of communication with a child in a child-care service robots scenario. For this application, we collect wearable sensors' data from the child and mother-child conversation data in daily life. Finally, we discuss our results and future works.

Introduction

As human-sized robots are popular and noticeable progress is evident in the human-robot interaction field, home-service robots have opened up a new market. Despite the significant developments related with home-service robots, a fundamental and inevitable problem is still an ongoing issue: to communicate with humans fluently and to generate more natural conversation considering the current context (Christensen, 2010).

Several researches have attempted to solve the problem mentioned above (Gorniak, 2004; Gorniak, 2005; Roy, 2005). These papers suggest consideration of the environmental information of robots in addition to speech information (spoken input, verbal input etc.) to generate conversation. For obtaining environmental information, these papers use perceived visual, auditory and spatial data from embedded sensors such as the camera and microphone.

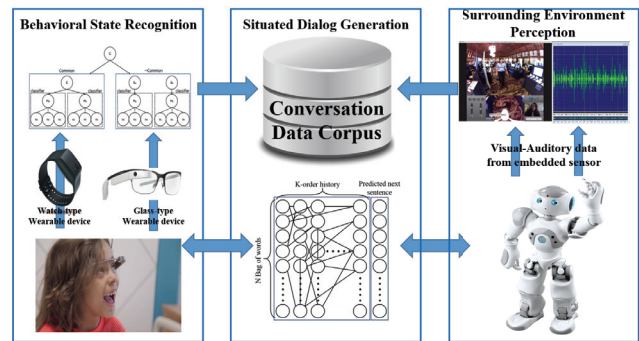


Figure 1: Situated conversation system architecture

However, these embedded sensors are quite limited to get the whole situational information on account of partial observability, which means that such sensors can only detect the environment surrounding the robot. Especially, to communicate with a human while considering the situation that the human is in, it needs to understand the current behavioral state of the human.

In this paper, we suggest using the behavioral signals of a child from wearable devices to gather information about the current situation of the human, in addition to using the robot's embedded sensors. With the sensor information, we can generate adapted conversations, which consider the sharing context together with the situation the human is in.

The system architecture is shown in Figure 1. The architecture consists of three parts: recognition of the behavioral state of the child, perception of the surrounding environment from the embedded sensor of robot and generation of situated conversation. As a preliminary research, we focused only on the wearable device for obtaining situational information, which is related to the first part of the system.

The remainder of the paper is organized as follows. In the next sections, we describe the algorithm for recognizing the user's behavioral state and generating situated con-



Figure 2: Private home studio environment





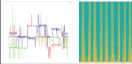
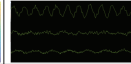
versation. After describing the child-care scenario as an application of this work, we conclude with a discussion of the future work and possible extensions.

Child-care Service Robot Scenario

Consider a scenario in which a home-service robot cares for a 10-year-old boy in his home on behalf of his mother. The robot should be able to wake him up in the morning, serve food for him, help him prepare for school, help him with his homework, play with him, etc. To carry out these tasks, the robot has to be able to talk with the child flexibly.

To simulate the situated conversation system involved in this scenario, we first changed the interior of the laboratory into a private house. This space consists of a living room, kitchen and the child’s room (Figure 2). We then recruited 2 pairs of participants, who represented the mother and child and collected wearable sensors’ data for the child and the mother-child natural conversation data in our laboratory. The conversation data is audio-recorded, and wearable sensors’ data is annotated for the child’s activity state, such as “Sleep” or “Meal”. The details of data specification are shown in table 1.

Table 1: Wearable devices and data specification

	Glass-type wearable device	Watch-type wearable device	Voice recorder
Equipment	 Google glass	 Empatica E4	 Sony IC Recorder
Original data	First-person video 32Hz	3-axis acceleration 32Hz	Speech signal 32Hz
Feature extractor	Convolutional Neural Network	Short-time Fourier Transform	Bag of words
Examples			
Wearable sensor labels	Activity in home (Meal, Sleep, Exercise, Dressing, Study, Wash)		

From these data, we can train our model to generate proper conversation. First, from the wearable sensors’ data, the Behavioral State Recognition Module is learned and from the conversation data corpus, we can train the sequential Hypernetwork model to generate the next utterance. Once the model is learned from the collected data, it can be used to interact with a child in natural language. The robot can infer the child’s behavioral state from the wearable sensors and can generate proper utterance based on the behavioral state and previous conversation.

Child’s Behavioral State Recognition

Wearable sensors are suited for analyzing the various response data of a person’s behavior in an unobtrusive way (Zhang, 2014). For example, a 3-axis accelerometer sensor on the wrist is used for analyzing movement of the body or physical behavior (Bruno, 2013). Also, the first-person’s video sensor from a glass-type wearable device is useful for recognizing the user’s activity (Doshi, 2015). However, sensor data are highly noisy because of external and internal factors. To reduce this affect, integrating the multi-modal sensors is very useful. Specifically, integrating the multi-modal sensors based on causal structure is highly recommended due to the fact that different sensors have different information quantity depending on their cause (Trommershauser, 2011).

Multi-modal Cue Integration Model

In this section, we suggest a machine learning model which recognizes a child’s behavioral state by efficiently integrating multi-modal sensor data. The structure of this model is shown in Figure 3. From sensor a and b, we can obtain the observed (raw) data o_a, o_b and process these to get more abstract information H_a and H_b , which are the estimated class labels from each sensor respectively by each of the sensor processing modules in this model. Therefore each sensor processing module is a classifier using only one sensor as an input. The variable S contains the true class label information which can be considered as a cause of the sensor data. The variable C indicates whether each sensor came from the common cause or a different cause. Now, it

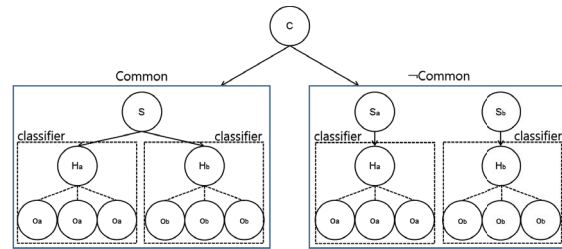


Figure 3: Behavioral state recognition module

is possible to calculate the probability of the common cause or different cause where the sensor's data came from and thus estimate the true class label based on this causal structure.

In detail, $P(S)$, $P(H_a|S)$, $P(H_b|S)$ follow categorical distribution:

$$P(S) = \text{Cat}(K, \mathbf{p}_S) = \prod_{i=1}^K p_i^{[i=S]} \quad (1)$$

$$P(H_a|S) = \text{Cat}(K, \mathbf{p}_{H_a|S}) = \prod_{i=1}^K p_i^{[i=H_a|S]} \quad (2)$$

$$P(H_b|S) = \text{Cat}(K, \mathbf{p}_{H_b|S}) = \prod_{i=1}^K p_i^{[i=H_b|S]} \quad (3)$$

The probability of common cause given H_a, H_b can be expressed by $P(C|H_a, H_b)$ and different cause can be expressed by $P(\neg C|H_a, H_b) = 1 - P(C|H_a, H_b)$. By applying Bayes rule, we can calculate

$$P(C|H_a, H_b) = \frac{P(H_a, H_b|C)P(C)}{P(H_a, H_b)} \quad (4)$$

$P(H_a, H_b)$ can be obtained by marginalization:

$$\mathbf{P}(\mathbf{H}_a, \mathbf{H}_b) \quad (5)$$

$$= \mathbf{P}(C)\mathbf{P}(\mathbf{H}_a, \mathbf{H}_b|C) + \mathbf{P}(\neg C)\mathbf{P}(\mathbf{H}_a, \mathbf{H}_b|\neg C)$$

Now, we can calculate $P(H_a, H_b|C)$ and $P(H_a, H_b|\neg C)$ as follows:

$$\begin{aligned} P(H_a, H_b|C) &= \\ & \sum_{i=1}^K P(H_a, H_b|s=i)P(s=i) \\ &= \sum_{i=1}^K P(H_a|s=i)P(H_b|s=i)P(s=i) \end{aligned} \quad (6)$$

$$\begin{aligned} P(H_a, H_b|\neg C) &= \\ & \left(\sum_{i=1}^K P(H_a|s_a=i)P(s_a=i) \right) \times \\ & \left(\sum_{j=1}^K P(H_b|s_b=j)P(s_b=j)(1 - \delta_{s_a, s_b}) \right) \end{aligned} \quad (7)$$

$$\text{where } \delta_{i,j} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

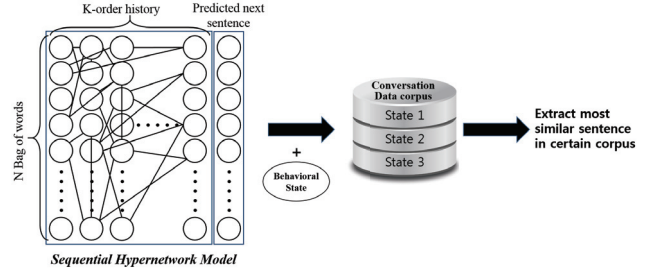


Figure 4: Situated conversation generation module

Because equation (5), (6) and (7) are discrete probability distributions, we can obtain $P(C|H_a, H_b)$ by calculating 4.

We can now estimate the real class label given the raw sensors' data as follows:

If $P(C|H_a, H_b) \geq 0.5$, the estimated class label \hat{S} given by raw data is

$$\begin{aligned} \hat{S} &= \underset{i}{\operatorname{argmax}} P(s=i|H_a, H_b) \\ &= \underset{i}{\operatorname{argmax}} P(H_a|s=i)P(H_b|s=i)P(s=i) \end{aligned} \quad (8)$$

If not, we compare between \hat{S}_a and \hat{S}_b and take the larger one as the estimated class label. \hat{S}_a and \hat{S}_b is expressed respectively by:

$$\hat{S}_a = \underset{i}{\operatorname{argmax}} P(s=i|H_a) \quad (9)$$

$$\hat{S}_b = \underset{i}{\operatorname{argmax}} P(s=i|H_b) \quad (10)$$

In this way, we can build a model which recognizes a child's behavioral state by using wearable sensors.

Situated Conversation Generation

Recently, most of the conversation generation models have used lots of prior knowledge, such as specific grammar for languages, to generate utterances (Shim, 2002; Reiter, 2000). However, their approaches are quite limited because it does not guarantee the scalability and the generality of the model. To tackle these problems, new approaches have been suggested containing a data-driven approach. In this section, we suggest a novel algorithm for generating the utterance considering the context and the situation based on the data-driven approach.

Situated Conversation Generation

As preliminary research to consider the situational information, we first distributed the Conversation data corpus according to their behavioral state labels: The behavioral

sub-corpus. Then, we designed a sequential Hypernetwork model to generate the next utterance in each behavioral sub-corpus. This model aims to model the probability distribution over the utterance $P(U_{future} | U_{past})$. To predict the probability of the next utterance, we model considering the previous k-step data $P(U_{t+1} | U_{t-k+1:t})$.

As the first step for learning the above model, the utterance input, which is an array of characters is transformed into a bag-of-words (BoW) vector. Then, the utterances in each behavioral sub-corpus are trained by the sequential Hypernetwork model (Zhang, 2008).

After training the sequential Hypernetwork with a conversation data corpus, the model can generate the next bag-of-words vector given the past sentence sequences and current behavioral state. The information of the current behavioral state is acquired by the child's behavioral state recognition module. Finally, the sentence which the robot must say/generate is determined by the model by choosing the most similar sentence with the new generated bag-of-words vector in the current behavioral sub-corpus.

Discussion

In this paper, we proposed a situated conversation system for a child-care service robot to communicate with a child fluently. To illustrate this system, we presented a child-care service robot scenario and our ongoing work. Our system consists of two parts: understanding the situational context and generating situated conversation. For considering the situational context, we used a child's behavioral state using wearable sensors, and generated proper utterance based on the situational context and previously held conversations.

As further work, we will integrate the three parts in one: perception of the surrounding environment by the robot's embedded sensors, recognition of a child's behavioral state by wearable sensors and generation of situated conversation by giving two information. To do this, we will also gather more real-world mother-child data in the future.

Acknowledgments

This work was partly supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (R0126-15-1072-SW.StarLab, 10044009-HRI.MESSI).

References

Bruno, B., Mastrogiovanni, F., Sgorbissa, A., Vernazza, T., and Zaccaria, R. 2013. Analysis of human behavior recognition algorithms based on acceleration data, Robotics and Automation (ICRA), 2013 IEEE International Conference on. IEEE.

Christensen, H., Kruijff, G. J. M., and Wyatt, J. (eds.), 2010. *Cognitive systems*. Vol. 8. Springer Science & Business Media.

Roy, D., and Reiter, E. 2005. Connecting language to the world. *Artificial Intelligence* 167.1: 1-12.

Doshi, J., Kira, Z., and Wagner, A. 2015. From Deep Learning to Episodic Memories: Creating Categories of Visual Experiences. *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*. ACS.

Gorniak, P., and Roy, D. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*: 429-470.

Gorniak, P., and Roy, D. 2005. Probabilistic grounding of situated speech using plan recognition and reference resolution. *Proceedings of the 7th international conference on Multimodal interfaces*. ACM.

Reiter, E., Dale, R., and Feng, Z. 2000. *Building natural language generation systems*. Vol. 33. Cambridge: Cambridge university press.

Shim, Y., and Kim, M. 2002. Automatic short story generator based on autonomous agents. *Intelligent Agents and Multi-Agent Systems*. Springer Berlin Heidelberg.

Trommershauser, J., Kording, K., and Landy, M. S. (eds.), 2011. *Sensory cue integration*. Oxford University Press.

Zhang, B.-T. 2008. Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory. *IEEE Computational Intelligence Magazine*, 3(3):49-63.

Zhang, B.-T. 2014. Ontogenesis of agency in machines: A multidisciplinary review. *AAAI 2014 Fall Symposium on The Nature of Humans and Machines: A Multidisciplinary Discourse*. AAAI press.