BELIEF NETWORK BASED DISAMBIGUATION OF OBJECT REFERENCE IN SPOKEN DIALOGUE SYSTEM FOR ROBOT

Yoko YAMAKATA   Tatsuya KAWAHARA   Hiroshi G. OKUNO

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{yamakata,kawahara,okuno}@kuis.kyoto-u.ac.jp

Abstract   We are studying joint activity in which a remote robot finds an object by communicating with the user over a voice-only channel. We focus on how the robot disambiguates the reference of the uttered word or phrase to the target object. For example, by "*cup*", one may refer to a "*teacup*", a "*coffee cup*", or even a "*glass*" under some situations. This reference (hereafter, "object reference") is user-dependent. We confirm that a user model of object references is significant by conducting a survey of 12 subjects. In addition to ambiguity of object reference, actual systems should cope with two other sources of uncertainty in speech and image recognition. We present a Belief Network based probabilistic reasoning system to determine the object reference. The resulting system demonstrates that the number of interactions needed to find a common reference is reduced as the user model is refined.

1. INTRODUCTION

Spoken natural language is the most important means of communication with robots. We are studying *joint activity* [1] in which a remote robot attempts to find a target by communicating with the user over a voice-only channel. As in Figure 1, the user knows about the area where the robot is, but is not able to see it. The user asks the robot to bring an object by saying, "*Could you bring me the cup on the table?*" The robot recognizes the utterance and looks for an appropriate object.

One of the most important problems in this study is how to disambiguate the reference of the uttered word or phrase to the target object. This reference (hereafter, "**object reference**") is user-dependent. Ambiguities in the object reference of the user's uttered word hinder smooth communications [2]. This also happens in human communication; for example, one may refer to the target object as a "*teacup*", while others may refer to the same object as a "*cup*" or even as a "*coffee cup*". If the user always designates only his favorite cup as the "*cup*" and the robot knows this fact, the robot should choose it immediately. However, if the user refers to any type of cup as a "*cup*", the robot cannot decide the target object without asking about other features such as
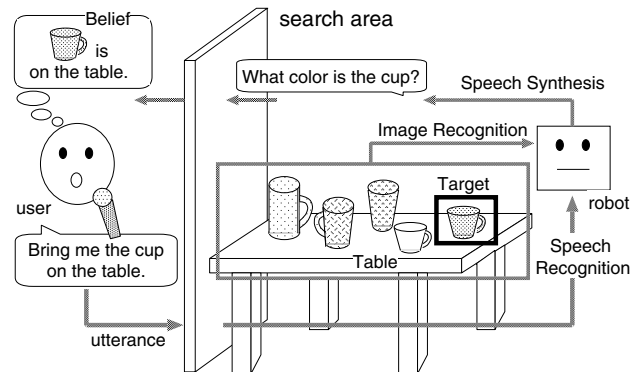

Figure 1. Joint activity of remote object search task

Name, *Pattern* and so on. On the contrary, the robot may find a target without disambiguating the user's utterances completely, if there is only one type of cup in the search area.

One of the conventional ways to cope with ambiguities of object reference is to assign a unique referent to each word. Most robots equipped with automatic speech recognition adopt this assumption of uniqueness, although it rarely holds true in the real world.

Another way to disambiguate the object reference is to allow user-initiative. For example, if the robot and the user share visual information of the environment, the user can always observe the robot to check whether it has misunderstood and correct it immediately [3, 4]. However, it is difficult for the robot to provide sufficient information to the user over a voice-only channel.

Since there are multiple sources of uncertainty such as in the automatic speech recognition and image recognition processes in addition to the object reference, we present a probabilistic reasoning method based on a **belief network** of the object reference. The robot selects an appropriate dialogue or a set of actions based on the belief network even under uncertainty from several sources. It revises the user model of object references by word when it finds a target, and maintains the user model for future dialogue.

image model / word				
glass	1	3	2	3
tumbler	3	1	3	3
teacup	1	3	1	1

Figure 2. A portion of the questionnaire. A subject is requested to answer whether each word-image mapping is acceptable (1), neutral (2), or unacceptable (3).

2. USER MODEL OF OBJECT REFERENCES

We conducted two surveys to confirm the twofold hypothesis that (i) there are significant inter-user differences in word reference and that (ii) individual users' reference remains stable for at least a month. The first survey is to obtain ambiguities in word-image mapping concerning various kinds of cups and glasses. We selected from a thesaurus 15 words that belong to the category of cup, e.g. "teacup", "glass", and "tumbler", and 14 images of line drawing illustrations. A portion of the questionnaire is depicted in Figure 2. Twelve subjects answered the questionnaire, that is, whether each word-image mapping is acceptable (1), neutral (2), or unacceptable (3). The second survey, conducted one month later, is to assess the personal variation with the same questionnaire.

As shown in Figure 3, the correlation between the responses in both surveys for each subject is obviously higher than the correlation between subjects in the first survey. The mean value of the correlation coefficient between the two surveys for the same subject is 0.77. This confirms that there is considerable variation in word-image mapping between users whereas individual users are consistent over time. The mean correlation between different subjects is 0.56. The result indicates that a personalized user model of word-image mapping represents actual user behavior better than a general model.

To acquire a user model of object references by word, we adopt incremental refinement. In other words, we start with an average user model as the initial model and adapt it incrementally by learning through the dialogue with the user.

3. BELIEF NETWORK FOR OBJECT REFERENCE

In addition with the user dependency, the resolution of object reference is hampered by errors in speech and image recognition errors. Noise in the real world and certain types of speech, especially utterances of elderly or children, degrade the speech recognition accuracy. The real world environment also influences image recognition. Since the robot inevitably operates in such environments, it must be robust against these ambiguities.
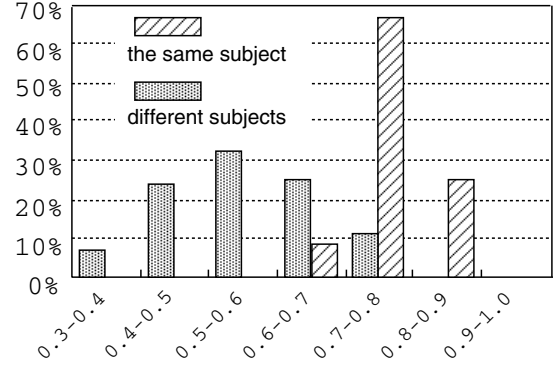

Figure 3. Correlation of the user models of object references between the same subjects in the two surveys and between different subjects in the first survey. X-axis indicates the range of correlation coefficient value.

In order to cope with these problems, we design a Belief Network (BN) that integrates speech-, language- and image-level information, and determine the objects in the search area by their belief factors obtained from the Belief Network.

Objects may be represented using attributes such as *Name*, *Color*, and *Pattern*. A specific BN should be constructed for each attribute. In the experiment we discuss below, only the BN for *Name* is implemented.

The BN as shown in Figure 4 has four levels; utterance, word, image model, and object. The three connections between adjacent levels correspond to the speech recognition, language understanding, and image recognition processes, respectively. In the example of the *Name* attribute as shown in Figure 4, keywords pertinent to the attribute are recognized from a user's utterance, and mapped to image models. These models are then matched with actual objects in the search area.

For each object, a belief factor measuring how well it matches the one intended by the user is computed. Here, suppose an utterance: X=*"Bring me the teacup."* The speech recognition process generates a set of keyword candidates and their confidence measures. Keywords are classified by their attribute (e.g. [Name [*teacup*, 0.7], [*cup*, 0.3]]). The confidence measure is obtained from the outputs of the speech recognition process: e.g.

$$Bel(teacup|X) = 0.7, Bel(cup|X) = 0.3.$$

The user model of object references is defined as a function $Assoc(word_w; model_m)$ from pairs of words and image models to degrees of association. The initial user model is derived from the results of the questionnaires described in section 2 by taking the mean values of all subjects for each item.

Finally, the belief factor that an object in the search area, $object_o$, is recognized as a model, $model_m$, is given by the similarity score of image processing as a confidence mea-
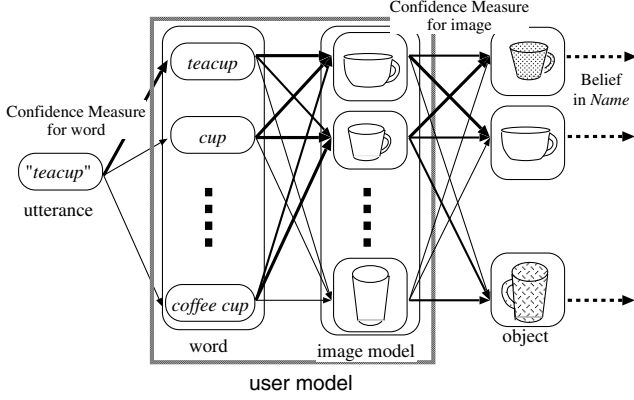
Figure 4. Belief Network for attribute *Name*

sure, $Bel(model_m | object_o)$. The belief factor that the utterance X refers to the object, $Bel(object_o | X)$, is calculated through the BN from the utterance level to the object level as follows:

$$Bel(model_m | X)$$
$$= \frac{\sum_p Assoc(word_p; model_m) Bel(word_p | X)}{\sum_p \sum_q Assoc(word_p; model_q) Bel(word_p | X)}$$

$$Bel(object_o | X)$$
$$= \sum_p Bel(model_p | object_o) Bel(model_p | X)$$

The degree of uncertainty in resolving the user's reference, $H(model_m | X)$, is defined as the entropy of belief factors of the image model layer in the BN:

$$H(model_m | X) = -\sum_p Bel(model_p | X) \log Bel(model_p | X)$$

When the degree of uncertainty is larger than a threshold, the robot asks the user about other attributes to reduce it.

The user may specify the object by multiple attributes (e.g. *Name* is "*teacup*" and *Color* is "*red*".). In this case, each belief factor is calculated separately, and then the combined belief factor, $Bel(object_o | ''teacup'', '' red'')$, is obtained by using Dempster-Shafer theory.

If there is a unique object in the search area whose score is maximal and exceeds the second-best score by 10 percent of maximal score, then it is chosen as the target. The system then updates the degree of association in the user model, $Assoc(word_w; model_m)$, for each attribute. In the example of "*teacup*" utterance X, the network of *Name* attribute is revised by adding the following difference:

$$\Delta Assoc(word_w; model_m)$$
$$= \eta \frac{Bel(word_w | ''teacup'')}{\sum_p Bel(word_p | ''teacup'')} \frac{Bel(model_m | object_t)}{\sum_q Bel(model_q | object_t)} E$$

where η is a learning rate and E is an increase of belief defined as below:

$$E = Bel(object_t | ''teacup'', '' red'') - Bel(object_t | ''teacup'')$$

Figure 5. Example of dialogue in the system

The dialogue system of the robot consists of two modules for resolving the object reference and planning dialogue, respectively. The object reference resolution module calculates the degree of uncertainty $H(model_m | X)$, and the dialogue planning module generates dialogue for disambiguation. If the reference cannot be resolved and the target is not determined regardless of all possible plans, the robot chooses one of the objects with maximum score (e.g. maximum $Bel(object_o | ''teacup'', '' red'')$) and proposes it to the user by rephrasing its object reference. In rephrasing name of the object, the dialogue planning module chooses the most suitable word for the object reference by tracing the BN inversely from the selected object $object_s$ to the word level. The belief factor, $Bel(word_w | object_s)$ is calculated as following:

$$Bel(word_w | object_s)$$
$$= \frac{\sum_p Bel(word_w | model_p) Bel(model_p | object_s)}{H(word_w)}$$

$$H(word_w)$$
$$= -\sum_q Bel(model_q | word_w) \log Bel(model_q | word_w)$$

The robot chooses the word of the largest belief factor and informs the user of the selected object with the expression.

4. EXPERIMENTS AND RESULTS

The system is implemented as a software robot in a virtual space using "Julian 3.1" automatic speech recognition and "CHATR version 0.94" speech synthesis engines on a SUN workstation. Image recognition is simulated by giving a confidence measure (between 0 and 1) for the image models.

We prepared two benchmark search areas as tasks for the experiment. The search areas have three local areas (kitchen, table, shelf), and there are 2, 3, or 4 objects in each local area, respectively. These objects are selected from seven types of cups, and there is more than one object of the same type within each benchmark test. In Benchmark 1 and 2, the same objects are used and their location areas are different. Eleven subjects were requested to search all objects with the robot by spoken dialogue. The subjects were

Table 1. Identification rate by planned dialogue

	Decision only by *Name*	Disambiguation by *Color*	Disambiguation by *Pattern*	Suggested Solution	Number of Dialogue Turns	Overall Achievement
Benchmark 1	19.2%	36.4%	2.0 %	23.2%	2.09	80.8%
Benchmark 2	40.4% (21.2% ↑)	22.2%	3.0%	17.2%	1.77	82.8%
Average	29.8%	29.3%	2.5%	20.2%	1.93	81.8%
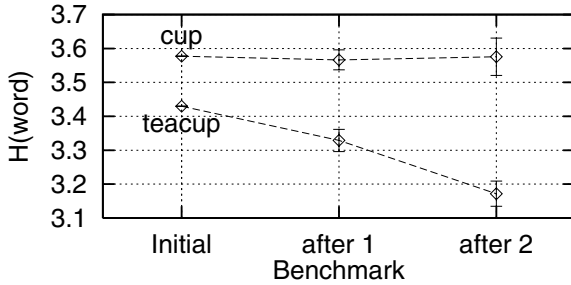
Figure 6. Transition of ambiguities $H(word)$ about "*cup*" and "*teacup*" of all subjects (maximum, average, and minimum value)

given pictures of the benchmark search areas beforehand, but could interact with the robot over a voice-channel only. They could choose any word for *Name* freely, but *Color* and *Pattern* were specified for each object. A dialogue fragment is shown in Figure 5.

The degree of uncertainty of the object reference for "*cup*" and "*teacup*" by the user model at the initial state, after benchmark 1, and after benchmark 2 is shown in Figure 6. According to the graph, "*cup*" is more ambiguous than "*teacup*" even at the initial state. Adapting the user model helps disambiguation for "*teacup*" but not for "*cup*" for most subjects. Since there are few image models referred to as "*teacup*" and they are typically fixed while there are many image models referred to as "*cup*", this result shows that the robot acquires the user models reasonably.

In addition, we measured to what extent the target is identifiable only by *Name* by adapting the user model. Table 1 shows the identification rate at a manually set threshold value of 3.4. At Benchmark 1, the robot could determine only 19.2% of all targets with failure rate of 2.0% due to the immature user model. At Benchmark 2 the user model is adequately acquired, and the identification rate is improved to 40.4% with 2.0% failure. The result confirms the effectiveness of the user model adaptation.

When the target is not identified, the robot generates a dialogue for disambiguation using the attributes of *Color* and *Pattern*, and, if necessary, proposes a solution by rephrasing the object name. Overall, the robot succeeded in finding 81.8% of all objects with an average of 1.93 dialogue turns. The main reasons of failures are that in some situations these two attributes are not sufficient to make a decision, and sometimes the user misunderstood the *Name* the robot used for specifying the target.

The robot changed the naming of objects as it adapted the user model. Initially the robot used six "*Name*" words for the objects in the search area. The number of names was reduced to 4.6 words by using the adapted user model as the subjects used 4.8 words on average. This shows that the robot adapts not only utterance-to-object but also object-to-utterance references. But in some cases, the user was found to adapt to the robot while the robot was attempting to adapt to the user. This phenomenon suggests the necessity to improve the belief revision system to follow the user's adaptation to the robot.

5. CONCLUSION

We have argued that the problem of object reference is essential in joint activity by spoken dialogue system. It is demonstrated that a general model of the object references by word concerning kinds of cups should be adapted to individual users. Based on this observation, we designed a system based on a Belief Network that integrates speech, language and image processing to disambiguate the object reference. It also adapts the understanding module by learning the user model in the framework of the Belief Network. The resulting system successfully reduces the ambiguity in identifying target objects.

6. REFERENCES

[1] Horvitz, E. and Paek, T.: A computational architecture for conversation. *Proceedings of Seventh International Conference on User Modeling*, Springer, 201–210, 1999.

[2] Cremers, A.: Object Reference in Task-Oriented Keyboard Dialogues. Harry Bunt (ed), *Multimodal Human-Computer Communication: systems, techniques, and experiments.*, Springer, 279–293, 1998.

[3] Inamura, T., Naka, K., Inaba, M., and Inoue, H.: Human-centered adaptive mobile robot based on on-line dialogue and stochastic experience representation. *Preprints of Fourth IFAC Symposium on Intelligent Autonomous Vehicle*, 61–66, 2001.

[4] Shinyama, Y., Tokunaga, T., and Tanaka, H.: Kairai — software robots understanding natural language. *3rd International Workshop on Human-Computer Conversation,* , Jul. 2000.