Learning to Mediate Perceptual Differences in Situated Human-Robot Dialogue

Changsong Liu and **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University
East Lansing, Michigan 48824
{cliu, jchai}@cse.msu.edu

Abstract

In human-robot dialogue, although a robot and its human partner are co-present in a shared environment, they have significantly mismatched perceptual capabilities (e.g., recognizing objects in the surroundings). When a shared perceptual basis is missing, it becomes difficult for the robot to identify referents in the physical world that are referred to by the human (i.e., a problem of referential grounding). To overcome this problem, we have developed an optimization based approach that allows the robot to detect and adapt to perceptual differences. Through online interaction with the human, the robot can learn a set of weights indicating how reliably/unreliably each dimension (e.g., object type, object color, etc.) of its perception of the environment maps to the human's linguistic descriptors and thus adjust its word models accordingly. Our empirical evaluation has shown that this weight-learning approach can successfully adjust the weights to reflect the robot's perceptual limitations. The learned weights, together with updated word models, can lead to a significant improvement for referential grounding in future dialogues.

Introduction

A new generation of robots have emerged in recent years to serve as humans' assistants and companions (Christensen, Kruijff, and Wyatt 2010). To allow humans to better interact with these robots, techniques to support situated human-robot dialogue become crucial. Unlike traditional spoken dialogue systems (McTear 2002) and conversational interfaces (Johnston et al. 2002; Chai et al. 2004), one significant challenge in situated human-robot dialogue is that the robot needs to perceive and make sense of the shared environment simultaneously during conversation. The robot's representation of the shared world is often limited by its preceptual (e.g., computer vision algorithms) and reasoning (i.e., inference algorithms) capabilities. Therefore, although co-present, humans and robots do not share a joint perceptual experience. The lack of a shared pereptual basis will jeopadize referential communication between the human and the robot. It will become more difficult for the robot to identify referents in the physical world that are referred to by the human, i.e., a problem of referential grounding.

Computational approaches to referential grounding often consist of two key components (Gorniak and Roy 2004; Siebert and Schlangen 2008; Liu, Fang, and Chai 2012). The first component addresses formalisms and methods that connect linguistic terms (e.g., *red*, *left*) to the lower level numerical features (e.g., rgb vectors) captured in the robot's representation of the perceived environment. We call this component the *word grounding models*. The second component extracts all the linguistic terms from a referring expression and combines their grounding models together to identify referents. For example, given a referring expression "*the big blue bottle to the left of the red box*", it will recognize that the intended referent has several attributes (e.g., color is blue, size is big, type is bottle) and it is to the left of another object. Then it will combine all relevant sensors' inputs and apply the corresponding word grounding models to identify the referent that most likely satisfies the referring expression.

Many previous works on referential grounding have focused on the first component, i.e., exploring how to learn and ground individual linguistic terms to low level physical attributes (e.g., Roy 2002; Barnard et al. 2003; Roy 2005). Although different algorithms have been applied for the second component (Liu, Fang, and Chai 2012; Liu et al. 2014), little attention has been paid to the question how to *intelligently* combine different attributes to ground references. However, this is an important question for human-robot dialogue since the human and the robot have mismatched representations of the shared environment. For example, the robot may not recognize any bottle, or may see something blue but not a bottle. Furthermore, the robot's perception of blue may be very different from the human's perception of blue. How should the robot utilize these different attributes? What part of its own perception should the robot trust the most when there is potential mismatch with the human's perception?

To address these questions, we have been investigating computational approaches that will allow the robot to learn and mediate perceptual differences during human-robot dialogue. The idea is that, by interacting with its human partner (e.g., through dialogue), the robot should be able to assess its perceptual differences from its human partner. In particular, the robot should be able to learn what dimension(s) of its own perception (e.g., recognition of objects or colors) are more reliable, namely more aligned with the human's perception reflected by the linguistic descriptions. To mediate the perceptual differences, the robot should use this self-assessment to update its internal models and further improve referential grounding in follow-up communication. Specifically, we have developed an optimization-based approach to efficiently learn a set of weights, which indicate how reliably/unreliably each dimension of the robot's perception of the environment maps to the human's linguistic descriptions. By simulating different types of mismatched perception, our empirical results have shown the weight-learning approach can successfully adjust the weights to reflect the robot's perceptual capabilities. In addition, the learned weights for specific linguistic terms (e.g., "red", "left") can further trigger automatic model updating for these words, which in turn leads to an improvement of referential grounding performance for future dialogues.

Related Work

Recent years have seen an increasing amount of work that grounds language to shared environment (Mooney 2008; Qu and Chai 2010; Chen and Mooney 2011; Tellex et al. 2011; Matuszek et al. 2012). Specifically, for grounding referring expressions to objects in the environment, different approaches have been developed (Gorniak and Roy 2004; 2007; Siebert and Schlangen 2008; DeVault and Stone 2009; Krishnamurthy and Kollar 2013). For example, Gorniak and Roy (2004) present an approach that grounds referring expressions to visual objects through semantic decomposition, using context free grammar that connects linguistic structures with underlying visual properties. To incorporate situational awareness, incremental approaches have been developed to prune interpretations which do not have corresponding visual referents in the environment (Scheutz et al. 2007; Kruijff et al. 2007). A recent work has developed an ontology to mediate different mental states between humans and robots (Lemaignan et al. 2012). In our previous work, we have incorporated collaborative discourse into referential grounding to mediate perceptual differences between humans and agents (Liu et al. 2013).

Most of these previous works have focused on how to connect the linguistic descriptions with the lower level physical attributes in referential grounding. In this paper, we address language grounding from a different angle. We focus on automatic assessing and adapting existing word grounding models by learning a set of informative weights from dialogues between humans and robots.

Method

Referential Grounding through Graph-Matching

We use *Attributed Relational Graph (ARG)* (Tsai and Fu 1979) to model the discourse of situated dialogue and the perceived environment. An ARG is a directed graph

$$G = (X, E) \tag{1}$$

in which

$X = \{x_m \mid m = 1, \dots, M\}$ is a set of M *nodes*;

$E = \{e_i = (x_t, x_h)_i \mid i = 1, \dots, I; x_t \in X; x_h \in X\}$ is a set of I *edges*, i.e., each edge is a pair of two nodes.

To store useful information, we further attach a set of *attributes* $\{v_a \mid a = 1, \dots, A\}$ to each node and each edge. The attributes of a node are used to represent object-specific properties, such as object-class (or type), color, and size. And the attributes of an edge are used to represent binary relations (e.g., spatial relations) between objects. For example, a node can be assigned a set of attributes as

$$\{v_1 = \text{'Apple'},\ v_2 = \text{'Big'},\ v_3 = \text{'Red'}\}$$

which specifies its type, size, and color.

Using the ARG representation, we can create a *dialogue graph* to represent the linguistic information gathered from the dialogue, and a *vision graph* to represent the perceived environment. The nodes in the dialogue graph correspond to the linguistic entities described in the dialogue, and the nodes in the vision graph correspond to the physical objects perceived from the environment. We use $G = (X, E)$ and $G' = (X', E')$ to denote the dialogue graph and the vision graph, respectively. A *matching* (denoted as Θ) between G and G' is to map each node in G to a corresponding node in G', formally defined as:

$$\Theta = \{(x_m, x'_n) \mid x_m \in X; x'_n \in X'\} \tag{2}$$

We further define the *compatibility function* of a matching Θ as

$$f(\Theta) = \sum_{x_m \in X} f_{\mathrm{X}}(x_m,\, x'_n) \;+\; \sum_{e_i \in E} f_{\mathrm{E}}(e_i,\, e'_j) \tag{3}$$

where x'_n/e'_j is the corresponding node/edge of x_m/e_i according to Θ. Then the optimal matching between G and G' is the one with the highest compatibility score:

$$\Theta^* = \arg\max_{\Theta} f(\Theta) \tag{4}$$

Unfortunately, finding the optimal matching between two graphs is a NP-hard problem. Thus we use a beam search algorithm (e.g., in (Cesar Jr et al. 2005)) to keep tractability.

Word Grounding Models

To compute $f(\Theta)$, we need to further define $f_{\mathrm{X}}(x_m,\, x'_n)$ and $f_{\mathrm{E}}(e_i,\, e'_j)$, i.e., the compatibility of a matched pair of nodes/edges. This is based on the attributes assigned to the pair of nodes/edges:

$$f_{\mathrm{X}}(x_m,\, x'_n) = \frac{1}{A} \sum_{a} \alpha_a\, f_a(v_a,\, v'_a) \tag{5}$$

$$f_{\mathrm{E}}(e_i,\, e'_j) = \frac{1}{B} \sum_{b} \beta_b\, f_b(v_b,\, v'_b) \tag{6}$$

in which $f_a(v_a,\, v'_a)$ and $f_b(v_b,\, v'_b)$ are what we call the "*word grounding models*" for the a-th node attribute and the b-th edge attribute, respectively. The input (v, v') is a pair of values on the same attribute, of which v is a linguistic value (i.e., a word) from the dialogue graph and v' is the corresponding visual feature from the vision graph. For example, some commonly used attributes and their typical values are shown in Table 1. The output of a word grounding model is

	Dialogue graph	*Vision graph*
Type	"apple"	recognized type, e.g., "ball"
Color	"red"	$(r : 210, g : 12, b : 90)$
Spatial relation	"left"	$(x : 300, y : 600)$

Table 1: Examples of commonly used attributes

a real number in the range of $[0, 1]$, which can be interpreted as a measurement of the compatibility between word v and visual feature v'.

As mentioned earlier, rather than learning word grounding models, the focus of this paper is to learn a set of weights to reflect the reliability of existing word grounding models (such as the learned models in (Roy 2002)). Thus, we define $\alpha_a \in [0, 1]$ and $\beta_b \in [0, 1]$ as the *weights* for the a-th node attribute and the b-th edge attribute, respectively. The weight of an attribute represents the reliability of the word grounding models associated with this attribute. For example, if object recognition is deemed not reliable, then a lower weight should be assigned to the *type* attribute.

Learning Weights through Optimization

The graph-matching algorithm relies on all these attributes to find out proper matchings and filter improper ones. Importantly, because the robot can have different capabilities in perceiving these different attributes, and linguistic referring expressions associated with different attributes can also have different levels of ambiguities, different attributes then should not always be treated equally. Thus we develop an optimization based approach to automatically acquire a set of weights based on the matching hypotheses and the ground-truth matching, which can be provided through confirmation sub-dialogue during real-time interaction.

The weights (i.e., α_a and β_b) are the "variables" that we aim to adjust, and our general objective is to maximize the reference grounding performance. We represent all the weights using a vector $\mathbf{w}$:

$$\mathbf{w} = [\alpha_1, \ldots, \alpha_A, \beta_1, \ldots, \beta_B]^T = [w_1, \ldots, w_K]^T \qquad (7)$$

i.e., $w_1 = \alpha_1, \ldots, w_A = \alpha_A, w_{A+1} = \beta_1, \ldots, w_K = \beta_B$ and $K = A + B$.

For a given matching Θ, its compatibility score $f(\Theta)$ then becomes a linear function of $\mathbf{w}$ as:

$$f(\Theta) = f_\Theta(\mathbf{w}) = \mathbf{P}_\Theta^T \mathbf{w} \qquad (8)$$

where

$$\mathbf{P}_\Theta = [P_1, \ldots, P_K]^T$$

is a vector of "coefficients" that are computed from the given Θ and predefined word grounding models.

Given two graphs G and G', suppose $\hat{\Theta}$ is the ground-truth matching, and $\Theta_1, \Theta_2, \ldots, \Theta_H$ are the top-H matching hypotheses (i.e., Θ_1 is the top-1 hypothesis and so forth) generated using an initial weights vector $\mathbf{w}_0$. If $\Theta_1 \neq \hat{\Theta}$, we can try to find a new $\mathbf{w}$ that may lead to a better matching outcome. This can be formulated into an optimization problem as:

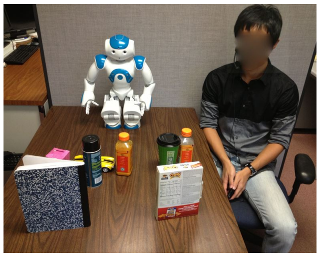


Figure 1: An example of the experiment setup

$$\begin{aligned}
&\max_{\mathbf{w},\varepsilon} \; \mathbf{P}_{\hat{\Theta}}^T \mathbf{w} - C \sum_{h=1}^{H} \varepsilon_h \\
&\text{s.t.} \\
&\qquad \mathbf{P}_{\Theta_1}^T \mathbf{w} - \mathbf{P}_{\hat{\Theta}}^T \mathbf{w} \le \varepsilon_1, \; \varepsilon_1 \ge 0 \\
&\qquad\qquad\qquad \vdots \\
&\qquad \mathbf{P}_{\Theta_H}^T \mathbf{w} - \mathbf{P}_{\hat{\Theta}}^T \mathbf{w} \le \varepsilon_H, \; \varepsilon_H \ge 0 \\
&\text{and} \\
&\qquad 0 \le w_k \le 1, \; \text{for all } k
\end{aligned} \qquad (9)$$

where $\{\varepsilon\}$ is a set of "slack" variables to accommodate infeasible constraints and C is a penalizing factor to prevent large values of ε. The essence of this optimization scheme is to find a proper $\mathbf{w}$, based on which the ground-truth matching can be ranked as high as possible among the matching hypotheses generated by the graph-matching algorithm. This now becomes a typical *linear programming* problem and can be solved efficiently using algorithms such as the Interior Point methods (Roos, Terlaky, and Vial 2006).

Empirical Evaluation

Data

To evaluate our weight-learning approach, we used the data collected from an earlier study that investigated collaborative efforts in human-robot dialogue (Chai et al. 2014). In that study, a human subject was instructed to play a set of object-naming games with a NAO robot, as shown in Figure 1. The human subject was given some secret names for several objects in the shared environment, and the task for the human was to communicate these secret names to the robot through dialogue. The NAO robot was equipped with fully automated components of computer vision, language understanding and dialogue management. We chose this object-naming task because it naturally led to dialogues focusing on referential communication (i.e., referring and grounding on the intended objects), as in the following example:

H1: The green cup is called Bill.
R1: OK, the green cup is Bill. (point to the inferred object)
H2: Do you see a car?
R2: Yes, I see a car.
H3: Do you see a blue can behind it?
R3: I don't see a blue can. But I see a blue unknown object there. Is that the can you are talking about? (point to the inferred object)
H4: Yes, it's called Mary.

	Type	Color	Spatial relation
Expressions	686	747	443

Table 2: Number of expressions of the most commonly used attributes in the data

R4: Got it. The blue can is Mary.

One controlled condition in the experiments was to simulate different perceptual capabilities of the robot, which resulted in two levels of variations:

- **High-Mismatch** simulated the situation where the human and the robot had a high mismatch in their perceptions of the shared environment. The robot's object-recognition error rate was manipulated to be very high, namely, a large portion (60% or 90%) of the objects were mis-recognized.[1]
- **Low-Mismatch** simulated the situation where the human and the robot had a low mismatch in their perceptions of the shared environment. The robot correctly recognized most of the objects, with only a small portion (10% or 30%) being mis-recognized.

Although the experiment was originally designed for a different purpose, it actually provides an ideal data set for evaluating our weight-learning approach. Since we currently do not address dialogue management, evaluating our algorithm only needs language inputs from the dialogue discourse and the corresponding visual environment. Furthermore, the systematic simulation of mismatched perceptions allows us to evaluate whether the weight-learning outcome is consistent with our expectation. For example, we would expect the learned weight for the attribute *type* to be smaller under the high-mismatch condition than under the low-mismatch condition. Before we present the results, we first briefly summarize the data.

There were a total of 147 dialogues collected from 24 human subjects. Among these dialogues, 73 were collected under the low-mismatch condition and 74 were under the high-mismatch condition. For each dialogue, the robot's perception of the environment, such as the object-recognition results, the color of each object (represented as a rgb vector) and the position of each object (represented as x and y coordinates), were also logged. Table 2 shows the number of referring expressions for three most frequently described attributes in the data. The *type* (e.g. "the bottle") and *color* (e.g. "the red object") attributes were the most commonly used to describe the referents. Besides, *spatial relation* (e.g. "the object in the front", "the bottle is to the right of the box") is also commonly used.

Weight-learning Results on Object-specific Attributes

We first applied the weight-learning approach on the two subsets of data (i.e., low-mismatch and high-mismatch) to see whether the expected weights can be learned. To make the evaluation more realistic, we further simulated an "online" learning scenario as the following: In each round of simulation, we randomly selected 20 dialogues as the training sequence. The weight-learning approach then went through these 20 dialogues one-by-one to learn and update the weights. More specifically, we first used the same language and vision processing system as in the data-collecting experiment (see (Chai et al. 2014) for more details) to build two graph representations, i.e., a language graph representing all the human subject's referring expressions and a vision graph representing the perceived environment. We then used the graph-matching algorithm to generate matching (i.e., referential grounding) hypotheses, and applied our weight-learning approach to learn a set of new weights[2]. With the learned new weights, we updated the current weights as:

$$w_t = w_{t-1} + \gamma(w_{new} - w_{t-1})$$

(γ is a step-size parameter set to be 0.5).

	Type	Color	Spatial relation
Low-mismatch	0.87	0.97	0.97
High-mismatch	0.45	0.9	0.71

Table 3: The final weights learned after 20 training dialogues (averaged over 100 runs of simulation)

We started with uniform weights (i.e., all being 1), and repeated the weight learning process throughout the sequence of the selected 20 training dialogues. Table 3 summarizes the final learned weights on the low-mismatch and high-mismatch data after going through the 20 training dialogues[3]. As we can see in Table 3, the most significant change from the low-mismatch condition to the high-mismatch condition is the drop of the learned weight for the "type" attribute (0.87 vs. 0.45). This is consistent with the situation (i.e., low-mismatch vs. high-mismatch) from which the data was collected. To further demonstrate the weight-learning efficiency, we plotted the updated weight of the "type" attribute after each training dialogue, as shown in Figure 2(a). It shows that when the robot's object-recognition was significantly mismatched with the human's perception (i.e, the high-mismatch condition), the weight for the "type" attribute quickly descended in the first 5 training dialogues, and after that it started to become stable and gradually converged to the final value.

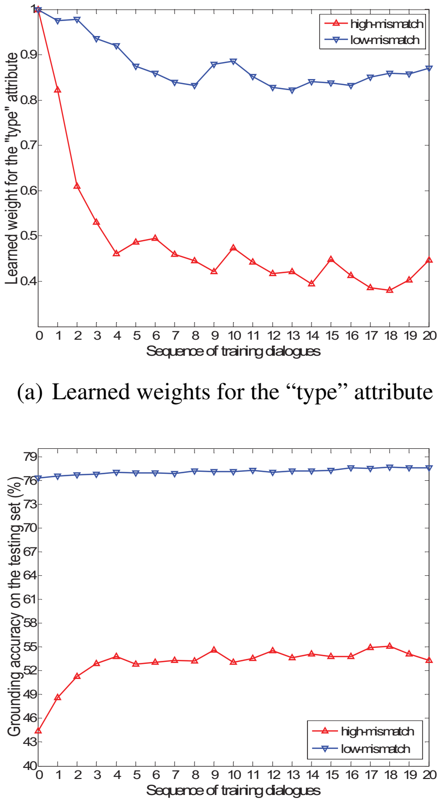
Besides the type attribute, the learned weights of the other attributes also indicate how reliable they are for referential grounding. The color attribute appears to be a reliable information source here, i.e., the color perception and grounding models are compatible with human descriptions. The spatial relation attribute is less reliable. This is possibly due to the vagueness of spatial expressions themselves, since a spatial expression such as "object in the front" can often result in several objects that all conform with the description and thus difficult to resolve based on the spatial information alone.

These learned weights not only indicate the robot's perceptual capabilities, but can also improve the referential grounding accuracy when applied to subsequent dialogues.

[1] These mis-recognized objects were randomly selected, and their object-recognition results were randomly assigned.

[2] To apply weight-learning, we assumed that the ground-truth matching of the current training dialogue is given.

[3] We have run the simulation 100 times, and the weights shown in Table 3 are the average over the 100 runs.



(a) Learned weights for the “type” attribute

(b) Referential grounding accuracies on the testing set by applying the learned weights

Figure 2: weight-learning and referential grounding results after each training dialogue (averaged over 100 runs)

To demonstrate this, we used all the remaining dialogues (i.e., those not selected as training dialogues) as the testing set[4]. After each training dialogue, we applied the current learned weights to generate referential grounding results on all the testing dialogues. The results (averaged referential grounding accuracies on the testing dialogues) are shown in Figure 2(b). Under the low-mismatch situation, applying the learned weights does not change the grounding accuracy. This is because the learned weights are close to the initial value (i.e., 1.0) as all the attributes were reasonably reliable. Under the high-mismatch situation, using the learned weights can improve grounding accuracy by 9.4% (from 44.4% to 53.8%) within the first 5 training dialogues. After that the grounding accuracy stays stable since the learned weights also become stable as shown in Figure 2(a).

Weight-learning Results at Word Level

Besides at the attribute level, it will be even more useful if the weights can be learned at a lower level, i.e., to learn a weight for each of the words that are used to describe an attribute. The learned weight then indicates the reliability of the robot’s perception and/or grounding model on that specific word. For example, if the robot can learn that its perception of “red” is unreliable, it can then adjust the grounding model for this specific word accordingly.

Color	orange	pink	green	blue
Without simulated errors	0.92	0.95	0.9	0.95
With simulated errors	0.47	0.59	0.88	0.97
Spatial relation	left	right	front	back
Without simulated errors	0.95	0.81	0.97	0.95
With simulated errors	0.19	0.73	0.77	0.73

Table 4: Final learned weights for some common words of the two attributes after 20 training dialogues (averaged over 100 runs of simulation)

To enable learning weights at the “word level”, instead of assigning only one weight for an attribute (i.e., all the words that describe one attribute always share the same weight), we need to assign each word grounding model a unique weight. The same weight learning approach can then be applied to learn how well the robot’s perception is aligned with the human’s description for each specific word. To evaluate word level weight-learning, we again used systematic simulation of perceptual errors which allowed us to easily assess whether expected weights can be learned given the simulated situation. Specifically, we used the low-mismatch data[5] and modified the robot’s perception to simulate some common errors that might happen in a real situation. The modifications we made were:

- For each object’s perceived color (i.e., an rgb vector), we increased the intensity of the r and g channels (by 100)[6] and decreased the intensity of the b channel (by 100). This was to simulate color-sensing error, e.g., due to environmental lighting noise or deficient color sensor.
- For each object’s perceived position (i.e., x and y coordinates), we decreased the x coordinate (by 300 pixels)[7]. This was to simulate spatial-sensing error, e.g., due to misaligned perspectives between the robot and the human or deficient spatial sensor.

With these simulated perceptual errors, we then used the same online learning scenario to evaluate the effectiveness and efficiency of weight-learning at the word level. Table 4 summarizes the final learned weights of some common words of two attributes after going through 20 randomly selected training dialogues. In the table we also show the learned weights from the situation that no errors were simulated, so that the weights learned with simulated errors can be compared. As we can see from Table 4, there are clear correspondences between the learned weights and the simulated errors in the robot’s perception:

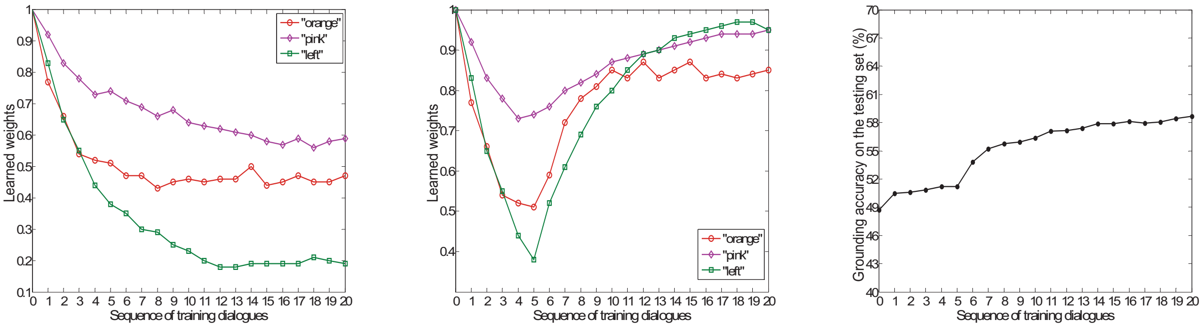
- For the “color” attribute, the learned weights indicate that the robot’s grounding of “orange” and “pink” is affected

[4] There are 53 and 54 testing dialogues for the low-mismatch and high-mismatch conditions, respectively.

[5] Since the low-mismatch data contained few original errors, it would be easier to see the effect of simulated errors here.

[6] The range of the intensity of each channel is from 0 to 255.

[7] The size of the image produced by the robot’s camera is 1280 × 960 pixels.



(a) Learned weights without model updating

(b) Learned weights with model updating

(c) Referential grounding accuracies with model updating

Figure 3: word level weight-learning evaluation results after each training dialogue (averaged over 100 runs of simulation)

by the simulated error of the inflated r and g channel intensities and the deflated b channel intensity.

- For the "spatial relation" attribute, the very low weight learned for "left" indicates the robot's problematic interpretation of this concept, which corresponds to the simulated error of shifting all the perceived objects to the left side.

These correspondences between the learned weights and the underlying perceptual errors again demonstrate that our weight-learning approach is capable of learning informative weights, which indicate how reliably the robot's perception maps onto the human's linguistic descriptions. For the efficiency of weight-learning at word level, Figure 3(a) shows the plots of updated weights for the words "orange", "pink", and "left", after each training dialogue during the online weight-learning process. It took only 5 training dialogues for the weight of "orange" to land on its final value. The weight of "pink" and "left" took some more training dialogues to reach the final value because they did not appear as often as the word "orange" in the data.

In a realistic setting, a more dynamic interaction and learning process would be expected. For example, the robot could adjust and improve its perceptual capabilities dynamically, based on the interaction with the human. Thus we further simulated one such dynamic process to see how our weight-learning responded to it. We still used the same data (i.e., low-mismatch data with simulated perceptual errors) and the online weight-learning process, but added a model-updating process after the first 5 training dialogues. This was to simulate the scenario that the robot automatically started to adjust its language grounding models for the unreliable words with low weights.

Initially, the grounding models for color and spatial relation terms were all defined as Gaussian distributions over the corresponding visual features (Roy 2002). To update the grounding models for the two color words (i.e., "orange" and "pink"), we followed the online word model acquisition approach as described in (Fang, Liu, and Chai 2012), which essentially kept updating the mean of the Gaussian distribution by averaging out the old mean and the new observed values. In addition to updating model parameters, the underlying models can be adjusted as well. For instance, for the word ”left", the robot can switch from the Gaussian model to the exponential decay function model as in (Regier and Carlson 2001).

Figure 3(b) shows the plots of learned weights for these three words with the model updating process. After adaptation, the weights for "orange", "pink", and "left" all become high values, indicating their grounding models became more consistent with the human's descriptions. We also evaluated referential grounding accuracy on the testing dialogues as we did earlier, but using both the updated models and weights after each training dialogue. As shown in Figure 3(c), referential grounding accuracy was improve by 10% (from 48.7% to 58.7%) with the updated models and weights, compared to the initial state of using the original models and uniform weights.

Conclusion

As a step towards enabling robust and adaptive human-robot dialogue, this paper presents a weight-learning approach for mediating the visual perceptual differences between a robot and its human partner in referential communication. As demonstrated by the empirical evaluation results, our weight-learning approach is capable of learning informative weights that reflect the alignment or misalignment between the robot's visual perception and the human's linguistic description of the shared environment. The learned weights can be applied to referential grounding and/or word model learning algorithms to improve the referential grounding performance. They can also be utilized by referring expression generation algorithms (e.g., Fang et al. 2013; Fang, Doering, and Chai 2014) to facilitate referential communication between robots and humans. Our current evaluation is based on several simplifications, including the simulation of perceptual errors and the strong assumption that the correct grounding information can be provided to the robot through dialogue with a human. In our future work, we will conduct experiments on a robot engaged in real-time dialogue with a human. Additionally, we will evaluate our algorithm with actual perceptual errors from real environments.

Acknowledgments

This work was supported by N00014-11-1-0410 from the Office of Naval Research and IIS-1208390 from the National Science Foundation. We thank anonymous reviewers for their valuable comments and suggestions.

References

Barnard, K.; Duygulu, P.; de Freitas, N.; Forsyth, D.; Blei, D.; and Jordan, M. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Cesar Jr, R. M.; Bengoetxea, E.; Bloch, I.; and Larrañaga, P. 2005. Inexact graph matching for model-based recognition: Evaluation and comparison of optimization algorithms. *Pattern Recognition* 38(11):2099–2113.

Chai, J.; Hong, P.; Zhou, M.; and Prasov, Z. 2004. Optimization in multimodal interpretation. In *Proceedings of ACL'04*, 1–8.

Chai, J. Y.; She, L.; Fang, R.; Ottarson, S.; Littley, C.; Liu, C.; and Hanson, K. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 33–40. ACM.

Chen, D. L., and Mooney, R. J. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 859–865.

Christensen, H. I.; Kruijff, G. M.; and Wyatt, J., eds. 2010. *Cognitive Systems*. Springer.

DeVault, D., and Stone, M. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 184–192.

Fang, R.; Liu, C.; She, L.; and Chai, J. Y. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of EMNLP'13*, 392–402.

Fang, R.; Doering, M.; and Chai, J. Y. 2014. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Fang, R.; Liu, C.; and Chai, J. Y. 2012. Integrating word acquisition and referential grounding towards physical world interaction. In *Proceedings of ICMI '12*, 109–116.

Gorniak, P., and Roy, D. 2004. Grounded semantic composition for visual scenes. In *Journal of Artificial Intelligence Research*, volume 21. 429–470.

Gorniak, P., and Roy, D. 2007. Situated language understanding as filtering perceived affordances. In *Cognitive Science*, volume 31(2). 197–231.

Johnston, M.; Bangalore, S.; Vasireddy, G.; Stent, A.; Ehlen, P.; Walker, M.; Whittaker, S.; and Maloor, P. 2002. MATCH: An architecture for multimodal dialogue systems. In *Proceedings of the ACL'02*, 376–383.

Krishnamurthy, J., and Kollar, T. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics* 1(2):193–206.

Kruijff, G.-J. M.; Lison, P.; Benjamin, T.; Jacobsson, H.; and Hawes, N. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symposium on Language and Robots*.

Lemaignan, S.; Ros, R.; Sisbot, E. A.; Alami, R.; and Beetz, M. 2012. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics* 4(2):181–199.

Liu, C.; Fang, R.; She, L.; and Chai, J. 2013. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, 78–86.

Liu, C.; She, L.; Fang, R.; and Chai, Y. J. 2014. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *Proceedings of ACL'14 (Volume 2: Short Papers)*, 13–18.

Liu, C.; Fang, R.; and Chai, J. Y. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the SIGDIAL 2012 Conference*, 140–149.

Matuszek, C.; FitzGerald, N.; Zettlemoyer, L.; Bo, L.; and Fox, D. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of ICML'12*.

McTear, M. F. 2002. Spoken dialogue technology: enabling the conversational interface. *ACM Computing Surveys* 34(1):90–169.

Mooney, R. J. 2008. Learning to connect language and perception. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 1598–1601.

Qu, S., and Chai, J. Y. 2010. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research* 37(1):247–278.

Regier, T., and Carlson, L. A. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General* 130(2):273.

Roos, C.; Terlaky, T.; and Vial, J.-P. 2006. *Interior point methods for linear optimization*. Springer.

Roy, D. K. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language* 16(3):353–385.

Roy, D. 2005. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences* 9(8):389–396.

Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Anderson, D. 2007. Incremental natural language processing for hri. In *Proceedings of the HRI'07*.

Siebert, A., and Schlangen, D. 2008. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of SIGDIAL'08*, 84–87.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S. J.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence*.

Tsai, W., and Fu, K. 1979. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *Systems, Man and Cybernetics, IEEE Transactions on* 9(12):757–768.