Towards Mediating Shared Perceptual Basis in Situated Dialogue

Changsong Liu, Rui Fang, Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48864
{cliu,fangrui,jchai}@cse.msu.edu

Abstract

To enable effective referential grounding in situated human robot dialogue, we have conducted an empirical study to investigate how conversation partners collaborate and mediate shared basis when they have mismatched visual perceptual capabilities. In particular, we have developed a graph-based representation to capture linguistic discourse and visual discourse, and applied inexact graph matching to ground references. Our empirical results have shown that, even when computer vision algorithms produce many errors (e.g. 84.7% of the objects in the environment are mis-recognized), our approach can still achieve 66% accuracy in referential grounding. These results demonstrate that, due to its error-tolerance nature, inexact graph matching provides a potential solution to mediate shared perceptual basis for referential grounding in situated interaction.

1 Introduction

To support natural interaction between a human and a robot, technology enabling human robot dialogue has become increasingly important. Human robot dialogue often involves objects and their identities in the environment. One critical problem is *interpretation and grounding of references* - a process to establish mutual understanding between conversation partners about intended references (Clark and Wilkes-Gibbs, 1986). The robot needs to identify referents in the environment that are specified by its human partner and the partner needs to recognize that the intended referents are correctly understood. It is critical for the robot and its partner to quickly and reliably reach the mutual acceptance of references before conversation can move forward.

Despite recent progress (Scheutz et al., 2007b; Foster et al., 2008; Skubic et al., 2004; Kruijff et al., 2007; Fransen et al., 2007), interpreting and grounding references remains a very challenging problem. In situated interaction, 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). Their knowledge and representation of the shared world are significantly different. When a shared perceptual basis is missing, grounding references to the environment will be difficult (Clark, 1996). Therefore, a foremost question is to understand how partners with mismatched perceptual capabilities mediate shared basis to achieve referential grounding.

To address this problem, we have conducted an empirical study to investigate how conversation partners collaborate and mediate shared basis when they have mismatched visual perceptual capabilities. In particular, we have developed a graph-based representation to capture linguistic discourse and visual discourse, and applied inexact graph matching to ground references. Our empirical results have shown that, even when the perception of the environment by computer vision algorithms has a high error rate (84.7% of the objects are mis-recognized), our approach can still correctly ground those mis-recognized objects with 66% accuracy. The results demonstrate that, due to its error-tolerance nature, inexact graph matching provides a potential solu-

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 140–149,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

tion to mediate shared perceptual basis for referential grounding in situated interaction.

In the following sections, we first describe an empirical study based on a virtual environment to examine how partners mediate their mismatched visual perceptual basis. We then provide details about our graph matching based approach and its evaluation.

2 Related Work

There has been an increasing number of published works on situated language understanding(Scheutz et al., 2007a; Foster et al., 2008; Skubic et al., 2004; Huwel and Wrede, 2006), focusing on interpretation of referents in a shared environment. Different approaches have been developed to resolve visual referents. Gorniak and Roy present an approach that grounds referring expressions to visual objects through semantic decomposition, using context free grammar that connect linguistic structures with underlying visual properties (Gorniak and Roy, 2004a). Recently, they have extended this work by including action-affordances (Gorniak and Roy, 2007). This line of work has mainly focused on grounding words to low-level 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., 2007a; Scheutz et al., 2007b; Brick and Scheutz, 2007). A recent work applies a bidirectional approach to connect bottom-up incremental language processing to top-down constrains on possible interpretation of referents given situation awareness (Kruijff et al., 2007). Most of these previous works address utterance level processing. Here, we are interested in exploring how the mismatched perceptual capabilities influences the collaborative discourse, and developing a graph-based framework for referential grounding with mismatched perceptions.

3 Empirical Study

It is very difficult to study the collaborative process between partners with mismatched perceptual capabilities. Subjects with truly mismatched perceptual capabilities are difficult to recruit, and the discrepancy between capabilities is difficult to measure and control. The wizard-of-oz studies with physical robots (e.g., as in (Green and Severinson Eklundh, 2001; Shiomi et al., 2007; Kahn et al., 2008)) are also insufficient since it is not clear what should be the underlying principles to guide the wizard's decisions and thus the perceived robot's behaviors (Steinfeld et al., 2009). To address these problems, motivated by the Map Task (Anderson et al., 1991) and the recent encouraging results from virtual simulation in Human Robot Interaction (HRI) studies (Carpin et al., 2007; Chernova et al., 2010), we conducted an empirical study based on virtual simulations of mismatched perceptual capabilities.

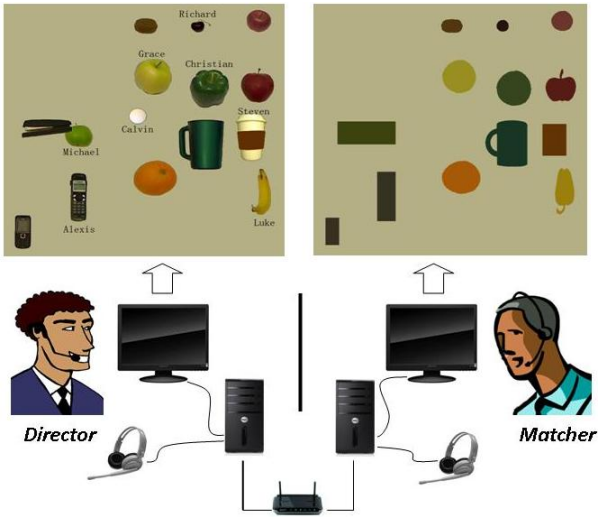

Figure 1: Our experimental system. Two partners collaborate on an object naming task using this system. The *director* on the left side is shown an (synthesized) original image, while the *matcher* on the right side is shown an impoverished version of the original image.

3.1 Experimental System and Task

The setup of our experimental system is shown in Figure 1. In the experiment, two human partners (a *director* and a *matcher*) collaborate on an object naming task. The mismatched perceptual capabilities between partners are simulated by different versions of an image shown to them: the director looks at an original image, while the matcher looks at an impoverished version of the original image.

The original image (the one on the left in Figure 1) was created by randomly selecting images of daily-life items (office supplies, fruits, etc.) from an image database and randomly positioning them onto a background. To create the impoverished im-

age (the one on the right in Figure 1), we applied standard Computer Vision (CV) algorithms to process the original image and then create an abstract representation based on the outputs from the CV algorithms.

More specifically, the original image was fed into a $segmentation \rightarrow feature\ extraction \rightarrow recognition$ pipeline of CV algorithms. First, the OTSU algorithm (Otsu, 1975) was used for image segmentation. Then visual features such as color and shape were extracted from the segmented regions (Zhang and Lu, 2002). Finally, object recognition was done by searching the nearest neighbor (in the shape-feature vector space) from a knowledge base of "known" objects. The impoverished image was then created based on the CV algorithms' outputs. For example, if an object in the original image was recognized as a pear, an abstract illustration of pear would be displayed in the impoverished image at the same position. Other features such like color and size of the object were also extracted from the original image and assigned to the illustration in the impoverished image.

In the naming task, the director's goal is to communicate the "secret names" of some randomly selected objects (i.e., target objects) in his/her image to the matcher, so that the matcher would know which object has what name. As shown in Figure 1, those secret names are displayed only on the director's screen but not the matcher's. Once the matcher believes that he/she correctly acquires the name of an target object, he/she will record the name by mouse-clicking on the target and repeating the name. A task is considered complete when the matcher has recorded the names of all the target objects.

3.2 Examples

Consistent with previous findings (Liu et al., 2011), our empirical study shows that human partners tend to combine object properties and spatial relations to construct their referring expressions. In addition, our empirical study has further demonstrated how partners manage to mediate their perceptual basis through collaborative discourse. Here are two examples from our data:

Example 1.

D[1]: the very top right hand corner, there is a red apple
M: ok
D: and then to the left of that red apple on the top of the screen is a red or black cherry
M: ok
D: and then to the left of that is a brown kiwi fruit
M: ok
D: and the, the red cherry is called Richard
. . . . . .

Example 2.

D: ok, um, so can we start in the top right
M: alright, um, the top right there are two rows of items, they are all circular or apple shaped
D: ok, um, the item in the very top right corner does not have a name
M: um, no name
M: um, to the left of that
D: yes, to the left of that is Richard
M: ok, are there only three items in that row
D: yes, there are only three
M: ok, this is Richard
. . . . . .

As shown in Example 1, the most commonly used object properties include *object class*, *color*, *spatial location*, and others such as *size*, *length* and *shape*. For the relations, the most common one is the projective spatial relations (Liu et al., 2010), such as *right*, *left*, *above*, *below*. Besides, as illustrated by Example 2, descriptions based on grouping of multiple objects are also commonly used. To mediate their shared basis, both the director and the matcher make extra effort to collaborate with each other. For instance, in Example 1, the director applies installment (Clark and Wilkes-Gibbs, 1986) where he utters noun phrases in episodes and the matcher explicitly accepts each installment before the director moves forward. In Example 2, the matcher intends to assist the grounding process by proactively providing what he perceives about the environment.

The data collected from our empirical study have indicated that, to mediate a shared perceptual basis and ground references, a successful method should consider the following issues: (1) It needs to capture the dynamics of the linguistic discourse and identify various relations among different referring expressions throughout discourse. (2) it needs to represent the perceived visual features and topological relations between visual objects in the visual discourse. (3) Because the perceived visual world by

[1] *D* stands for *Director* and *M* for *Matcher*.

the matcher (who represents the lower-calibre artificial agent) very often differs from the perceived visual world by the director (who represents the higher-calibre human partner), reference resolution will need some approximation without enforcing a complete satisfaction of constraints. Based on these considerations, we have developed a graph-based approach for referential grounding. Next we give a detailed account on this approach.

4 A Graph-based Approach to Referential Grounding

In the field of image analysis and pattern recognition, Attributed Relational Graph (ARG) is a very useful data structure to represent an image (Tsai and Fu, 1979; Sanfeliu and Fu, 1983). In an ARG, the underlying unlabeled graph represents the topological structure of the scene. Then each node and edge are labeled with a vector of attributes that represents local features of a single node or the topological features between two nodes. Based on the ARG representations, an inexact graph matching is to find a graph or a subgraph whose *error-transformation cost* with the already given graph is minimum (Eshera and Fu, 1984).

Motivated by the representation power of ARG and the error-correcting capability of inexact graph matching, we developed a graph-based approach to address the referential grounding problem. ARG and probabilistic graph matching have been previously applied in multimodal reference resolution (Chai et al., 2004a; Chai et al., 2004b) by integrating speech and gestures. Here, although we use similar ARG representations, our algorithm is based on inexact graph matching and our focus is on mediating shared perceptual basis.

4.1 Graph Representations

Figure 2 illustrates the key elements and the process of our graph-based method. The key elements of our method are two ARG representations, one of which is called the *discourse graph* and the other called the *vision graph*.

The discourse graph captures the information extracted from the linguistic discourse.[2] To create the discourse graph, the linguistic discourse first needs

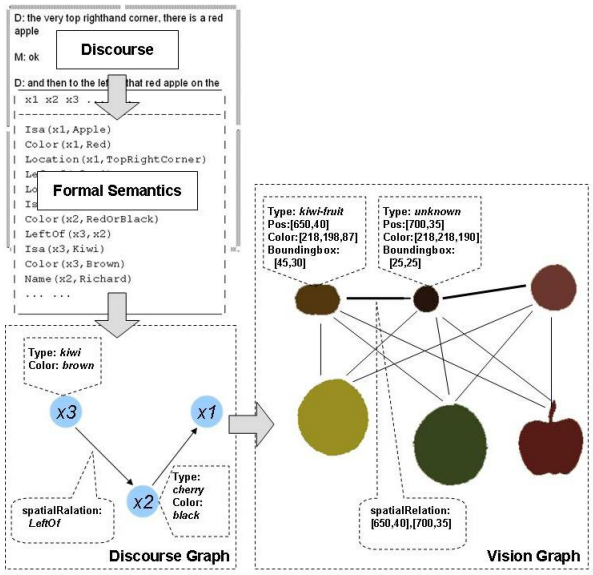

Figure 2: An illustration of graph representations in our method. The discourse graph is created from formal semantic representations of the linguistic discourse; The vision graph is created by applying CV algorithms on the corresponding scene. Given the two graphs, referential grounding is to construct a node-to-node mapping from the discourse graph to the vision graph.

to be processed by NLP components, such as the semantic composition and discourse coreference resolution components. The output of the NLP components are usually in the form of some formal semantic representations, e.g. in the form of first-order logic representations. The discourse graph is then created based on the formal semantics, i.e. each new discourse entity corresponds to a node in the graph, one-arity predicates correspond to node attributes and two-arity predicates correspond to edge attributes. The vision graph, on the other hand, is a representation of the visual features extracted from the scene. Each object detected by CV algorithms is represented as a node in the vision graph, and the attributes of the node correspond to visual features, such as the color, size and position of the object. The edges between nodes represent their relations in the physical space.

Given the discourse graph and the vision graph, now we can formulate referential grounding as constructing a node-to-node mapping from the discourse graph to the vision graph, or in other words, a matching between the two graphs. Note that, the

[2] Currently we only focus on the utterances from the director.

matching we encounter here is different from the original graph matching problem that is often used in the image analysis field. The original version only considers matching between two graphs that have the same type of values for each attribute. But in the case of referential grounding, all the attributes in the discourse graph possess symbolic values since they come from formal semantic representations, whereas the attributes in the vision graph are often numeric values produced by CV algorithms. Our solution is to introduce a set of *symbol grounding functions*, which bridges the heterogeneous attributes of the two graphs and makes general graph matching algorithms applicable to referential grounding.

4.2 Inexact Graph Matching

We formulate referential grounding as a graph matching problem, which has extended the original graph matching approach used in image processing and pattern recognition filed (Tsai and Fu, 1979; Tsai and Fu, 1983; Eshera and Fu, 1984).

First, we give the formal definition of an ARG, which is a doublet of the form

$$G = (N, E)$$

where

N	The set of attributed-nodes of graph G, defined as $N = \{(i, a) \,\vert 1 \leq i \leq \vert N \vert\}$.
E	The set of directed attributed-edges of graph G, defined as $E = \{(i, j, e) \,\vert 1 \leq i, j \leq \vert N \vert\}$.
$(i, a) \in N$	Node i with a as its attribute vector, where $a = [v_1, v_2, \cdots, v_K]$ is a vector of K attributes. To simplify the notation, We will denote a node as a_i.
$(i, j, e) \in E$	The directed edge from node i to node j with e as its attribute vector, where $e = [u_1, u_2, \cdots, u_L]$ is a vector of L attributes. We will denote an edge as e_{ij}.

In an ARG, the value of a node/edge attribute v_k/u_l can be symbolic, numeric, or as a vector of numeric values. For example, if v_1 is used to represent the color feature of an object, then a possible assignment could be $v_1 = [255, 0, 0]$, which is the rgb color vector.

Suppose we represent referring expressions from the linguistic discourse as a discourse graph G and objects perceived from the environment as a vision graph G', referential grounding then becomes a graph matching problem: given $G = (N, E)$ and $G' = (N', E')$, in which

$$N = \{a_i \,\vert 1 \leq i \leq I\}, E = \{e_{i_1 i_2} \,\vert 1 \leq i_1, i_2 \leq I\}$$

$$N' = \{{a_j}' \,\vert 1 \leq j \leq J\}, E' = \{{e'}_{j_1 j_2} \,\vert 1 \leq j_1, j_2 \leq J\}$$

A matching between G and G' is to find a one-to-one mapping between the nodes in N and the nodes in N'.

Note that it is not necessary for every node in N or N' to be mapped to a corresponding node in the other graph. If a node is not to be mapped to any node in the other graph, we describe it as being mapped to Λ, which denotes an abstract "null" node. To represent the matching result, we re-order N and N' such that the first I'/J' nodes in N/N' are those which have been mapped to their corresponding nodes in the other graph, and the nodes after them are the unmatched nodes, i.e. those matched with Λ. Then the matching result is

$$\begin{aligned} M &= M_1 \cup M_2 \cup M_3 \\ &= \{(i, j) \,\vert 1 \leq i \leq I', 1 \leq j \leq J'\} \\ &\quad \cup \{i \,\vert I' < i \leq I\} \\ &\quad \cup \{j \,\vert J' < j \leq J\} \end{aligned}$$

Here M_1 is a set of I' pairs of indices of matched nodes. M_2 and M_3 are the sets of indices of all the unmatched nodes in N and N', respectively. Then M is what we call a matching between G and G'. It is an inexact matching in the sense that we allow both G and G' to have a subset of nodes, i.e. M_2 and M_3, that are not matched with any node in the other graph (Conte et al., 2004). The cost of a matching M is then defined as

$$C(M) = C(M_1) + C(M_2) + C(M_3)$$

To complete the definition of $C(M)$, we use M_{11} to denote the set of all the first indices of the matched pairs in M_1, i.e. $M_{11} = \{i \,\vert 1 \leq i \leq I'\}$, and $H = \left(N^H, E^H\right)$ is the subgraph of G that is induced by the subset of nodes $N^H = \{a_i \,\vert i \in M_{11}\}$, then we have

$$\begin{aligned} C(M_1) &= \sum_{a_i \in N^H} C_N(a_i, {a'}_j) + \\ &\quad \sum_{e_{i_1 i_2} \in E^H} C_E(e_{i_1 i_2}, {e'}_{j_1 j_2}) \end{aligned}$$

$$\begin{aligned} C(M_2) &= \sum_{a_i \in (N - N^H)} C_N(a_i, \Lambda) + \\ &\quad \sum_{e_{i_1 i_2} \in (E - E^H)} C_E(e_{i_1 i_2}, \Lambda) \end{aligned}$$

in which $C_N(a_i, {a'}_j)$ is the cost of mapping a_i to ${a'}_j$, $C_E(e_{i_1 i_2}, {e'}_{j_1 j_2})$ is the cost of mapping $e_{i_1 i_2}$ to ${e'}_{j_1 j_2}$, and $C_N(a_i, \Lambda)$/$C_E(e_{i_1 i_2}, \Lambda)$ is the cost of mapping a_i/$e_{i_1 i_2}$ to the null node/edge. They are also called node/edge substitution cost and node/edge insertion cost, respectively (Eshera and Fu, 1984). Note that, in our case we let $C(M_3) = 0$ since we have assumed that the size of G' is bigger than the size of G.

Finally, the optimal matching between G and G' is the one with the minimum matching cost

$$M^* = \arg\min_M \ C(M)$$

which gives us the most feasible result of grounding the entities in the discourse graph with the objects in the vision graph.

Given our formulation of referential grounding as a graph matching problem, the next question is how to find the optimal matching between two graphs. Unfortunately, such a problem belongs to the class of *NP-complete* (Conte et al., 2004). In practice, techniques such as A^* search are commonly used to improve the efficiency, e.g. in (Tsai and Fu, 1979; Tsai and Fu, 1983). But the memory requirement can still be considerably large if the heuristic does not provide a close estimate of the future matching cost (Conte et al., 2004). In our current approach, we use a simple beam search algorithm (Zhang, 1999) to retain the tractability. Following the assumption in (Eshera and Fu, 1984), we set the beam size as hJ^2, where h is the current level of the search tree and J is the size of the bigger graph (in our case G').

4.3 Symbol Grounding Functions

As mentioned in Section 4.1, in referential grounding the discourse graph and the vision graph possess different types of attribute values, therefore we introduce a set of "symbol grounding functions", based on which node/edge substitution and insertion costs can be formally defined.

We start with node substitution cost to give a formal definition of symbol grounding functions. As defined in the previous section, the node substitution cost of mapping (substituting) node a with node a' is[3]

$$C_N(a, a')$$

Recall that in our definition of ARG, each node is represented by a vector of attributes, i.e. $a = [v_1, v_2, \cdots, v_K]$ and $a' = [{v'}_1, {v'}_2, \cdots, {v'}_K]$. Thus, we define the node substitution cost as

$$C_N(a, a') = \sum_{k=1}^{K} -\ln \ f_k(v_k, {v'}_k)$$

in which $f_k(v_k, {v'}_k) = p \ (p \in [0, 1])$ is what we call the symbol grounding function for the k-th attribute.

More specifically, a symbol grounding function for the k-th attribute takes two input arguments, namely v_k and ${v'}_k$, which are the values of the k-th attribute from node a and a' respectively. The output of the function is a real number p in the range of $[0, 1]$, which can be interpreted as a measurement of the compatibility between a symbol (or word) v_k and a visual feature value ${v'}_k$.

Let $\mathcal{L} = \{w_1, w_2, \cdots, w_Z, \textit{UNK}\}$ be the set of all possible symbolic values of v_k, then $f_k(v_k, {v'}_k)$ can be further decomposed as

$$f_k(v_k, {v'}_k) = \begin{cases} f_{k1}({v'}_k) & \text{if } v_k = w_1; \\ f_{k2}({v'}_k) & \text{if } v_k = w_2; \\ \vdots & \vdots \\ f_{kZ}({v'}_k) & \text{if } v_k = w_Z; \\ \lambda_k & \text{if } v_k = \textit{UNK}. \end{cases}$$

Here the idea is that each value of v_k may specify an unique function that determines the compatibility of a visual feature value ${v'}_k$. For example, suppose that we are defining a symbol grounding function for the attribute of "spatial location", i.e. where is an object located in the environment. The symbolic value v can be in the set of $\{\textit{Top}, \textit{Bottom}, \cdots, \textit{UNK}\}$, and the visual feature value v' is the x and y coordinates (in pixels) of the object's center of mass in the image. A grounding function for the symbol *Top* can be defined as[4]

$$f_{\textit{Top}}(v') = f_{\textit{Top}}(x, y) = \begin{cases} 1 - \frac{y}{800} & \text{if } y < 400; \\ 0 & \text{otherwise.} \end{cases}$$

Note that we have added a special symbol *UNK* to represent the "unknown" (or "unspecified") value of v_k. When the value of an attribute in the discourse graph is unknown, i.e. the speaker did not mention anything about a particular property, the grounding function will simply return a predefined

[3] For the ease of notation we have dropped the subscript of a node.

[4] Assume that the size of the image is 800×800 pixels and the left-top corner is the origin $(0, 0)$

Type of Error	Number of Objects
No Error	9 (5.1%)
Recognition Error	150 (84.7%)
Segmentation Error	18 (10.2%)
Total	177

Table 1: Types of errors among all the target (named) objects. *Recognition error*: an object is incorrectly recognized as another type of object, or an unknown type. *Segmentation error*: an object is missing, or merged with another object.

Type of Error	Accuracy/Detection Rate	
	Object-properties Only	Object-properties and Relations
No Error	66.7% (6/9)	77.8% (7/9)
Recognition Error	38.7% (58/150)	66% (99/150)
Segmentation Error	33.3% (6/18)	44.4% (8/18)
Overall	39.5% (70/177)	64.4% (114/177)

Table 2: Referential grounding performance of our method. The accuracy/detection rates in the table were obtained by comparing the results with annotated ground truths.

constant, which we denote as λ. The node insertion cost $C_N(a, \Lambda)$ is now defined as[5]

$$C_N(a, \Lambda) = \sum_{k=1}^{K} -\ln\ \lambda_k$$

Currently we set all the symbol grounding functions' outputs for the unknown value (i.e. the λs) to ε, which is an arbitrarily small real number ($\varepsilon > 0$).

5 Empirical Results

Three pairs of subjects participated in our experiment. Each pair (one acted as the director and the other as the matcher) completed the naming task on 8 randomly created images. In total we collected 24 dialogues with 177 target objects to be named. Table 1 summarizes the errors made by the CV algorithms when the 177 named objects from the original images were processed and represented in the impoverished images, as described in Section 3.1. As shown in the table, only 5% of the objects were correctly represented in the impoverished images. The other 95% of objects were either mis-recognized (about 85%) or mis-segmented (10%).

The evaluation of our approach is based on whether the target objects are correctly grounded by the graph matching method. To focus our current effort on the referential grounding aspect, we ignored all the matchers' contributions to the dialogues. Thus the discourse graphs were built based on only the director's utterances. The formal semantics of each of the director's valid utterances was manually annotated using the DRS (Discourse Representation Structure) representation (Bird et al., 2009). The discourse graphs were then generated from the annotated formal semantics. The vision graphs were generated from the outputs of the CV algorithms. The graph matching method was then applied to return a (sub-) optimal matching between the two graphs.

Table 2 shows the referential grounding performance of our method. To better understand the advantages of the graph-based approach, we have compared two settings. In the first setting, only the object-specific properties are considered for computing the comparability between a linguistic expression and a visual object, and the relations between objects are ignored. This setting is similar to the baseline approach used in (Prasov and Chai, 2008; Prasov and Chai, 2010). In the second setting, the complete graph-based approach is applied, i.e. both the object's properties and the relations between objects are considered. As shown in Table 2, although the improvements of performance for the no-error objects and mis-segmented objects are not significant due to the small sample sizes, the performance for the mis-recognized objects is significantly improved by 27.3% ($p < .001$). The improvement for the overall performance is also significant (by 24.9%, $p < .001$). The comparison between two settings have demonstrated the importance of representing and reasoning on relations between objects in referential grounding, and the graph-based approach provides an ideal solution to capture relations.

In particular, even CV error rate is high (due to the simple CV algorithms we used), our method is still able to achieve 66% accuracy of grounding the mis-recognized objects. Furthermore, when a referred object is completely "missing" in the vision graph

[5]The edge substitution/insertion cost is defined in the same way as the node substitution/insertion cost.

due to segmentation error[6], our method is capable to detect such discrepancy between linguistic input and visual perception. The results have shown that 44.4% of those cases have been correctly detected. This is also a very important aspect since information about failures of grounding will allow the dialogue manager and/or the vision system to adapt better strategies.

6 Discussions

The work presented here only represents an initial step in our on-going investigation towards mediating shared perceptual basis in situated dialogue. It consists of several simplifications which will be addressed in our future work.

First, the discourse graph is created only based on contributions from the director, using manual annotations of formal semantics of the discourse. As shown in the examples (Section 3.2), the collaborative discourse has rich dynamics reflecting participants' collaborative behaviors. So our future work is to model these different discourse dynamics and take them into account in the creation of the discourse graph. The discourse graph will be created after each contribution as the conversation unfolds. When utterances are automatically processed, semantics of these utterances often will not be extracted correctly or completely as in their manual annotations. Therefore, our future work will also explore how to efficiently match hypothesized discourse graphs (from automated semantic processing) with vision graphs.

Second, our current symbol grounding functions are very simple and intuitive. Our future work will explore more sophisticated models that have theoretical motivations (e.g., grounding spatial terms based on the Attentional Vector Sum (AVS) model (Regier and Carlson, 2001)) and enable automated acquisition of these functions (Roy, 2002; Gorniak and Roy, 2004b). In addition, we will explore context-based symbol grounding functions where context will be explicitly modeled. Grounding a linguistic term to a visual feature will be influenced by contextual factors such as surroundings of the environment, the discourse history, the speaker's individual preference, and so on.

Lastly, as shown in our examples, the matcher also contributes significantly to ground references. This appears to suggest that, in situated dialogue, lower-calibre partners (i.e., robot, and here the matcher) also make extra effort to ground references. The underlying motivation could be their urge to match what they perceive from the environment to what they are told by their higher-calibre partners (i.e., human). This motivation can be potentially modeled as graph-matching and can be used to guide the design of system responses. We will explore this idea in the future.

7 Conclusion

In situated human robot dialogue, a robot and its human partners have significantly mismatched capabilities in perceiving the environment, which makes grounding of references in the environment especially difficult. To address this challenge, this paper describes an empirical study investigating how human partners mediate the mismatched perceptual basis. Based on this data, we developed a graph-based approach and formulate referential grounding as inexact graph matching. Although our current investigation has several simplifications, our initial empirical results have shown the potential of this approach in mediating shared perceptual basis in situated dialogue.

Acknowledgments

This work was supported by Award #1050004 and Award #0957039 from National Science Foundation and Award #N00014-11-1-0410 from Office of Naval Research.

References

A.H. Anderson, M. Bader, E.G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python*. O'Reilly Media.

T. Brick and M. Scheutz. 2007. Incremental natural language processing for hri. In *Proceeding of the*

[6]For example, if the director refers to "a white ball" but CV algorithm fails to detect that object from the environment, then the node in the discourse graph representing "a white ball" should not be mapped to anything in the vision graph.

ACM/IEEE international conference on Human-Robot Interaction (HRI-07), pages 263–270.

S. Carpin, M. Lewis, J. Wang, S. Balakirsky, and C. Scrapper. 2007. USARSim: a robot simulator for research and education. In *Proceedings of the 2007 IEEE Conference on Robotics and Automation*.

J.Y. Chai, P. Hong, and M.X. Zhou. 2004a. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 70–77. ACM.

J.Y. Chai, P. Hong, M.X. Zhou, and Z. Prasov. 2004b. Optimization in multimodal interpretation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 1. Association for Computational Linguistics.

S. Chernova, J. Orkin, and C. Breazeal. 2010. Crowdsourcing hri through online multiplayer games. *AAAI Symposium on Dialogue with Robots*.

H. H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. In *Cognition*, number 22, pages 1–39.

H. H. Clark. 1996. *Using language*. Cambridge University Press, Cambridge, UK.

D. Conte, P. Foggia, C. Sansone, and M. Vento. 2004. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(3):265–298.

M. A. Eshera and K. S. Fu. 1984. A graph distance measure for image analysis. *IEEE transactions on systems, man, and cybernetics*, 14(3):398–410.

M.E. Foster, E.G. Bard, R.L. Hill, M. Guhe, J. Oberlander, and A. Knoll. 2008. Generating haptic- ostensive referring expressions in cooperative, task-based human-robot dialogue. *Proceedings of ACM/IEEE Human-Robot Interaction*.

B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. 2007. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of HRI07*, pages 73–80.

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

P. Gorniak and D. Roy. 2004b. Grounded semantic composition for visual scenes. *J. Artif. Intell. Res. (JAIR)*, 21:429–470.

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

A. Green and K. Severinson Eklundh. 2001. Task-oriented dialogue for CERO: a user centered approach. In *Proceedings of 10th IEEE international workshop on robot and human interactive communication*, September.

Sonja Huwel and Britta Wrede. 2006. Situated speech understanding for robust multi-modal human-robot communication. In *Proceedings of the International Conference on Computational Linguistics (COLING/ACL)*.

P. Kahn, N. Greier, T. Kanda, H. Ishiguro, J. Ruckert, R. Severson, and S. Kane. 2008. Design patterns for sociality in human-robot interaction. In *Proceedings of HRI*, pages 97–104.

Geert-Jan M. Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, and Nick Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symposium on Language and Robots*.

C. Liu, J. Walker, and J.Y. Chai. 2010. Disambiguating frames of reference for spatial language understanding in situated dialogue. In *AAAI Fall Symposium on Dialogue with Robots*.

C. Liu, D. Kay, and J.Y. Chai. 2011. Awareness of partners eye gaze in situated referential grounding: An empirical study. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

N. Otsu. 1975. A threshold selection method from gray-level histograms. *Automatica*, 11:285–296.

Z. Prasov and J.Y. Chai. 2008. What's in a gaze?: the role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29. ACM.

Z. Prasov and J.Y. Chai. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 471–481. Association for Computational Linguistics.

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

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

A. Sanfeliu and K. S. Fu. 1983. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, 13(3):353–362.

M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson. 2007a. First steps toward natural human-like HRI. In *Autonomous Robots*, volume 22.

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

M. Shiomi, T. Kanda, S. Koizumi, H. Ishiguro, and N. hagita. 2007. Group attention control for communication robots with wizard of OZ approach. In *Proceedings of HRI*, pages 121–128.

M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock. 2004. Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, 34(2):154–167.

A. Steinfeld, O. C. Jenkins, and B. Scassellati. 2009. The oz of wizard: Simulating the human for interaction research. In *Proceedings of HRI*, pages 101–107.

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

W.H. Tsai and K.S. Fu. 1983. Subgraph error-correcting isomorphisms for syntactic pattern. *year: 1983*, 13:48–62.

D. Zhang and G. Lu. 2002. An integrated approach to shape based image retrieval. In *Proc. of 5th Asian conference on computer vision (ACCV)*, pages 652–657.

W. Zhang. 1999. *State-space search: Algorithms, complexity, extensions, and applications*. Springer-Verlag New York Inc.