Towards Situated Dialogue: Revisiting Referring Expression Generation

Rui Fang, Changsong Liu, Lanbo She, Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, 48824, USA
{fangrui,cliu,shelanbo,jchai}@cse.msu.edu

Abstract

In situated dialogue, humans and agents have mismatched capabilities of perceiving the shared environment. Their representations of the shared world are misaligned. Thus referring expression generation (REG) will need to take this discrepancy into consideration. To address this issue, we developed a hypergraph-based approach to account for group-based spatial relations and uncertainties in perceiving the environment. Our empirical results have shown that this approach outperforms a previous graph-based approach with an absolute gain of 9%. However, while these graph-based approaches perform effectively when the agent has perfect knowledge or perception of the environment (e.g., 84%), they perform rather poorly when the agent has imperfect perception of the environment (e.g., 45%). This big performance gap calls for new solutions to REG that can mediate a shared perceptual basis in situated dialogue.

1 Introduction

Situated human robot dialogue has received increasing attention in recent years. In situated dialogue, robots/artificial agents and their human partners are co-present in a shared physical world. Robots need to automatically perceive and make inference of the shared environment. Due to its limited perceptual and reasoning capabilities, the robot's representation of the shared world is often incomplete, error-prone, and significantly mismatched from that of its human partner's. Although physically co-present, a joint perceptual basis between the human and the robot cannot be established (Clark and Brennan, 1991). Thus, referential communication between the human and the robot becomes difficult.

How this mismatched perceptual basis affects referential communication in situated dialogue was investigated in our previous work (Liu et al., 2012). In that work, the main focus is on reference resolution: given referential descriptions from human partners, how to identify referents in the environment even though the robot only has imperfect perception of the environment. Since robots need to collaborate with human partners to establish a joint perceptual basis, referring expression generation (REG) becomes an equally important problem in situated dialogue. Robots have much lower perceptual capabilities of the environment than humans. How can a robot effectively generate referential descriptions about the environment so that its human partner can understand which objects are being referred to?

There has been a tremendous amount of work on referring expression generation in the last two decades (Dale, 1995; Krahmer and Deemter, 2012). However, most existing REG algorithms were developed and evaluated under the assumption that agents and humans have access to the same kind of domain information. For example, many experimental setups (Gatt et al., 2007; Viethen and Dale, 2008; Golland et al., 2010; Striegnitz et al., 2012) were developed based on a visual world for which the internal representation is assumed to be known and can be represented symbolically. However, this assumption no longer holds in situated dialogue with robots. There are two important distinctions in situated dialogue. *First, the perfect knowledge of the environment is not available to the agent ahead of time.* The agent needs to automatically make inferences to connect recognized lower-level visual features with

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 392–402,
Seattle, Washington, USA, 18-21 October 2013.

symbolic labels or descriptors. Both recognition and inference are error-prone and full of uncertainties. *Second, in situated dialogue the agent and the human have mismatched representations of the environment.* The agent needs to take this difference into consideration to identify the most reliable features for REG. Given these two distinctions, it is not clear whether state-of-the-art REG approaches are applicable under mismatched perceptual basis in situated dialogue.

To address this issue, this paper revisits the problem of REG in the context of mismatched perceptual basis. We extended a well known graph-based approach (Krahmer et al., 2003) that has shown to be effective in previous work (Gatt and Belz, 2008; Gatt et al., 2009). We incorporated uncertainties in perception into cost functions. We further extended regular graph representation into hypergraph representation to account for group-based spatial relations that are important for visual descriptions (Dhande, 2003; Tenbrink and Moratz, 2003; Funakoshi et al., 2006; Liu et al., 2012). Our empirical results demonstrate that both enhancements lead to about a 9% absolute performance gain compared to the original approach. However, while our approache performs effectively when the agent has perfect knowledge or perception of the environment (e.g., 84%), it performs poorly under the mismatched perceptual basis (e.g., 45%). This performance gap calls for new solutions for REG that are capable of mediating mismatched perceptual basis.

In the following sections, we first describe our hypergraph-based representations and illustrate how uncertainties from automated perception can be incorporated. We then describe an empirical study using Amazon Mechanical Turks for evaluating generated referring expressions. Finally we present evaluation results and discuss potential future directions.

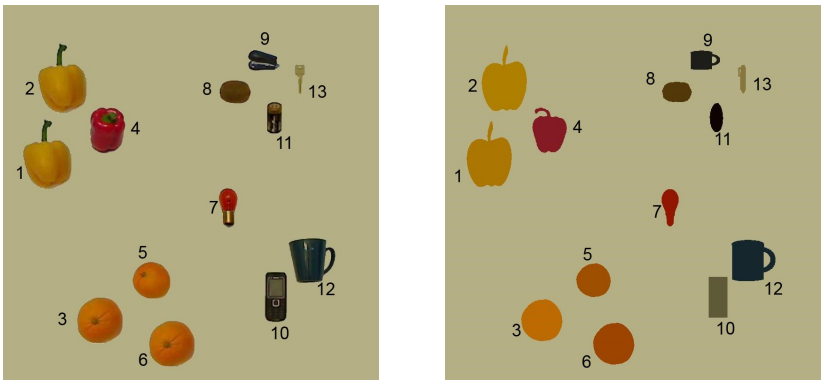
2 Related Work

Since the Full Brevity algorithm (Dale, 1989), many approaches have been developed and evaluated for REG (Dale, 1995; Krahmer and Deemter, 2012), such as the incremental algorithm (Dale, 1995), the locative algorithm (Kelleher and Kruijff, 2006), and graph-based approaches (Krahmer et al., 2003; Croitoru and Van Deemter, 2007). Most of these approaches assume the agent has access to a complete symbolic representation of the domain. While these approaches work well for many applications involving user interfaces, the question is whether they can be extended to the situation where the agent has incomplete or incorrect knowledge and needs to make inference about the domain or the world.

Recently, there has been increasing interest in REG for visual objects (Roy, 2002; Golland et al., 2010; Mitchell et al., 2013). Some work (Golland et al., 2010) uses visual scenes that are generated by computer graphics and thus the internal representation of the scene is known. Some other work focuses on the connection between lower-level visual features and symbolic descriptors for REG (Roy, 2002; Mitchell et al., 2013). However, most work assumes no vision recognition errors. It is well established that automated recognition of visual scenes is extremely challenging. This process is error-prone and full of uncertainties. It is not clear whether the existing approaches can be extended to the situation where the agent has imperfect perception of the shared environment.

An earlier work by Horacek (Horacek, 2005) has looked into the problem of mismatched knowledge between conversation partners for REG. The approach is a direct extension of the incremental algorithm (Dale, 1995). However, this work only provides a proof of concept example to illustrate the idea. No empirical evaluation was given.

All these previous works have motivated our present investigation. We are interested in REG under mismatched perceptual basis between conversation partners, where the agent has imperfect perception and knowledge of the shared environment. In particular, we took a well-studied graph-based approach (Krahmer et al., 2003) and extended it to incorporate group spatial relations and uncertainties associated with automated perception of the environment. The reason we chose a graph-based approach is that graph representations are widely used in the fields of computer vision (CV) and pattern recognition to represent spatially rich scenes. Nevertheless, the findings from this investigation provide insight to other approaches.



(a) an original scene

(b) the corresponding impoverished scene

Figure 1: An original scene and its impoverished scene processed by CV algorithm

3 Hypergraph-based REG

Towards mediating a shared perceptual basis in situated dialogue, our previous work (Liu et al., 2012) has conducted experiments to study referential communication between partners with mismatched perceptual capabilities. We simulated mismatched capabilities by making an original scene (Figure 1(a)) available to a director (simulating higher perceptual calibre) and a corresponding impoverished scene (Figure 1(b)) available to a matcher (simulating lowered perceptual calibre). The impoverished scene is created by re-rendering automated recognition results of the original scene by a CV algorithm. An example of the original scene and an impoverished scene is shown in Figure 1. Using this setup, the director and the matcher were instructed to collaborate with each other on some naming games. Through these games, they collected data on how partners with mismatched perceptual capabilities collaborate to ground their referential communication.

The setup in (Liu et al., 2012) is intended to simulate situated dialogue between a human (like the director) and a robot (like the matcher). The robot has a significantly lowered ability in perception and reasoning. The robot's internal representation of the shared world will be much like the impoverished scene which contains many recognition errors. The data from (Liu et al., 2012; Liu et al., 2013) shows that different strategies were used by conversation partners to produce referential descriptions. Besides directly describing attributes or binary relations with a relatum, they often use group-based descriptions (e.g., *a cluster of four objects on the right*). This is mainly due to the fact that some objects are simply not recognizable to the matcher. Binary spatial relationships sometimes are difficult to describe the target object, so the matcher must resort to group information to distinguish the target object from the rest of the objects. For example, suppose the matcher needs to describe the target object 5 in Figure 1(b), he/she may have to start by indicating the group of three objects at the bottom and then specify the relationship (i.e., top) of the target object within this group.

The importance of group descriptions has been shown not only here, but also in previous works on REG (Funakoshi et al., 2004; Funakoshi et al., 2006; Weijers, 2011). While the original graph-based approach can effectively represent attributes and binary relations between objects (Krahmer et al., 2003), it is insufficient to capture within-group or between-group relations. Therefore, to address the low perceptual capabilities of artificial agents, we introduce hypergraphs to represent the shared environment. Our approach has two unique characteristics compared to previous graph-based approaches: (1) A hypergraph representation is more general than a regular graph. Besides attributes and binary relations, it can also represent group-based relations. (2) Unlike previous work, here the generation of hypergraphs are completely driven by automated perception of the environment. This is done by incorporating uncertainties in perception and reasoning into cost functions associated with graphs. Next we

give a detailed account on hypergraph representation, cost functions incorporating uncertainties, and the search algorithm for REG.

3.1 Hypergraph Representation

A directed hypergraph G (Gallo et al., 1993) is a tuple of the form: $G = \langle X, A \rangle$, in which

$$X = \{x_m\}$$
$$A = \{a_i = (t_i, h_i) \mid t_i \subseteq X, h_i \subseteq X\}$$

Similar to regular graphs, a hypergraph consists of a set of nodes X and a set of arcs A. However, different from regular graphs, each arc in A is considered as a *hyperarc* in the sense that it can capture relations between any two subsets of nodes: a tail (t_i) and a head (h_i). Therefore, a hypergraph is a generalization of a regular graph. It becomes a regular graph if the cardinalities of both the tail and the head are restricted to one for all hyperarcs. While regular graphs are commonly used to represent binary relations between two nodes, hypergraphs provide a more general representation for n-ary relations among multiple nodes.

We use hypergraphs to represent the agent's perceived physical environment (also called *scene hypergraphs*). More specifically, each perceived object is represented by a node in the graph. Each perceived visual attribute of an object (e.g., color, size, type information) or a group of objects (e.g., number of objects in the group, location) is captured by a self-looping hyperarc. Hyperarcs are also used to capture the spatial relations between any two subsets of nodes, whether it is a relation between two objects, or between two groups of objects, or between one or more objects within a group of objects.

For example, Figure 2 shows a hypergraph created for part of the impoverished scene shown in Figure 1(b) (i.e., the upper right corner including objects 7, 8, 9, 11, and 13). One important characteristic is that, because the graph is created based on an automated vision recognition system, the values of an attribute or a relation in the hypergraph are numeric (except for the type attribute). For example, the value of the *color* attribute is the RGB distribution extracted from the corresponding visual object, the value of the *size* attribute is the width and height of the bounding box and the value of the *location* attribute is a function of spatial coordinates. These numerical features will be further converted to symbolic labels with certain confidence scores (described later in Section 3.3.2).

3.2 Hypergraph Pruning

The perceived visual scene can be represented as a complete hypergraph, in which any pair of two subsets of nodes are connected by a hyperarc. However, such a complete hypergraph is not only inefficient but also unnecessary. Instead of keeping all possible n-ary relations (i.e., hyperarcs), we only retain those relations that are likely used by humans to produce referring expressions, based on two heuristics.

The first heuristic is based on perceptual principles, also called the Gestalt Laws of perception (Sternberg, 2003), which describe how people group visually similar objects into entities or groups. Two well known principles of perceptual grouping are proximity and similarity (Wertheimer, 1938): objects that lie close together are often perceived as groups; objects of similar shape, size or color are more likely to form groups than objects differing along these dimensions. Based on these two principles, previous works have developed different algorithms for perceptual grouping (Thrisson, 1994; Gatt, 2006). In our investigation, we adopted Gatt's algorithm (Gatt, 2006), which has shown to be more accurate for spatial grouping. Given the results from spatial grouping, we only retain hyperarcs that represent spatial relations between two objects, between two perceived groups, between one object and a perceived group, or between one object and the group it belongs to.

The second heuristic is based on the observation that, given a certain orientation, people tend to use a relatum that is closer to the referent than more distant relata. In other words, it is less likely to refer to an object relative to a distant relatum when there is a closer relatum. For example, when referring to the stapler (object 9 in Figure 1(a)), it is more likely to use "the stapler above the battery" than "the stapler above the cellphone". Based on this observation, we prune the hypergraphs by only retaining hyperarcs between an object and their closest relata for each possible orientation.

Figure 2 shows the resulting hypergraph for representing a subset of objects (7, 8, 9, 11, and 13) in Figure 1(a).

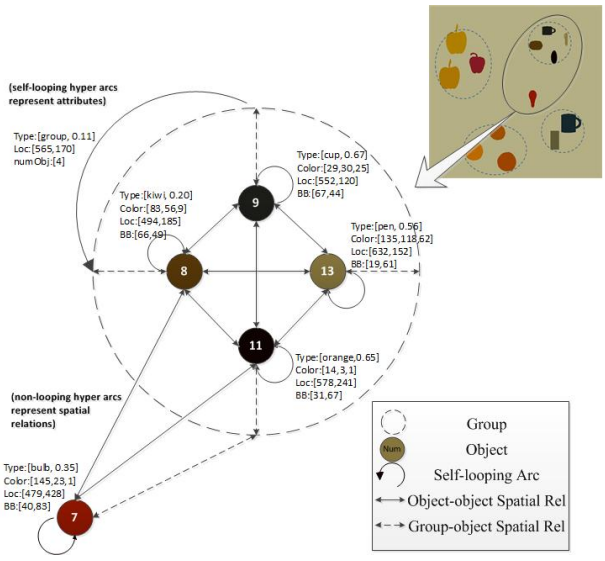


Figure 2: An example of hypergraph representing the perceived scene (a partial scene only including object 7, 8, 9, 11, 13 for Figure 1(a)).

3.3 Symbolic Descriptors for Attributes

As mentioned earlier, the values of attributes of objects and their relations are numerical in nature. In order for the agent to generate natural language descriptions, the first step is to assign symbolic labels or descriptors to those attributes and relations. Next we describe how we use a lexicon with grounded semantics in this process.

3.3.1 Lexicon with Grounded Semantics

Grounded semantics provides a bridge to connect symbolic labels or words with lower level visual features (Harnad, 1990). Previous work has developed various approaches for grounded semantics mainly for the reference resolution task, i.e., identifying visual objects in the environment given language descriptions (Dhande, 2003; Gorniak and Roy, 2004; Tenbrink and Moratz, 2003; Siebert and Schlangen, 2008; Liu et al., 2012). For the referring expression generation task here, we also need a lexicon with grounded semantics.

In our lexicon, the semantics of each category of words is defined by a set of semantic grounding functions that are parameterized on visual features. For example, for the *color* category it is defined as a multivariate Gaussian distribution based on the RGB distribution. Specific words such as *green, red, or blue* have different means and co-variances as the following:

$$color : red = f_r(\vec{v}_{color}) = N(\vec{v}_{color} \mid \mu_1, \sum_1)$$
$$color : green = f_g(\vec{v}_{color}) = N(\vec{v}_{color} \mid \mu_2, \sum_2)$$
$$color : blue = f_b(\vec{v}_{color}) = N(\vec{v}_{color} \mid \mu_3, \sum_3)$$

The above functions define how likely a set of recognized visual features (i.e., $\vec{v}_{color}$) describing the color dimensions (i.e., RGB distribution) is to match the color terms *red*, *green*, and *blue*.

For the spatial relation terms such as *above, below, left, right*, the semantic grounding functions take both vertical and horizontal coordinates of two objects, as follows [1]:

$$spatialRel : above(a, b) = f_{above}(\vec{v_{aloc}}, \vec{v_{bloc}})$$
$$= \begin{cases} 1 - \frac{|x_a - x_b|}{400} & \text{if } y_a < y_b; \\ 0 & \text{otherwise.} \end{cases}$$

Using the above convention, we have defined semantic grounding functions for *size* category words (e.g., *small* and *big*) and absolute position words (e.g., *top*, *below*, *left*, and *right*). In addition, we use object recognition models (Zhang and Lu, 2002) to define *class type* category words such as *apple* and *orange* used in our domain.

3.3.2 Attribute Descriptors and Cost Functions

Given the lexicon with grounded semantics as described above, the numerical attributes captured in the scene hypergraph can be converted to symbolic descriptors. For each attribute (e.g., *color*) or relation, the corresponding visual feature vector (i.e., $\vec{v}_{color}$) is plugged into the semantic grounding functions for the corresponding category of words. The word that best describes the attribute is chosen as the descriptor for that attribute. For example, given an RGB color distribution $\vec{v}_{color}$, we can find the color descriptor as follows:

$$color : w^* = \underset{red,green,blue}{\arg\max} \; f_w(\vec{v}_{color}),$$

For each attribute or relation, we can find a best descriptor in this manner. In addition, we also obtain a numerical value (returned from the semantic

[1] The size of the overall scene is 800x800.

grounding functions) that measures how well this descriptor describes the corresponding visual features. Intuitively, one would choose a descriptor that closely matches the visual features. Based on this intuition, we define the cost for each attribute A as the following:

$$cost(A) = 1 - f_{w^*}(\vec{v_A})$$

where w^* is the best descriptor for the attribute.

Given an attribute, the better the descriptor matches the extracted visual features, the lower the cost of the corresponding hyperarc.

3.4 Graph Matching for REG

Now the hypergraph representing the perceived environment has symbolic descriptors for its attributes and relations together with corresponding costs. Given this representation, REG can be formulated as a graph matching algorithm similar to that described in (Krahmer et al., 2003). We use the same Branch and Bound algorithm described in (Krahmer et al., 2003). In this approach, a hypothesis hypergraph (starting with one node representing the target object) is gradually expanded by adding in a least cost hyperarc from the scene hypergraph. At each expansion, the hypothesis graph is matched against the scene hypergraph to decide whether it matches any nodes other than the target node in the scene hypergraph. The expansion stops if the hypothesis graph does not cover any other nodes except for the target node. At this point, the hypothesis graph captures all the content (e.g., attributes and relations) required to uniquely describe the target object. We then apply a set of simple generation templates to generate the surface form of referring expressions based on the hypothesis graph.

4 Empirical Evaluations

4.1 Evaluation Setup

To evaluate the performance of this hypergraph-based approach to REG, we conducted a comparative study using crowd-sourcing. More specifically, we created 48 different scenes similar to that in Figure 1(a). Each scene has 13 objects on average and there are 621 objects in total. For each of these scenes, we applied a CV algorithm (Zhang and Lu, 2002) and generated scene hypergraphs as described in Section 3.1. We then use different generation strategies (varied in terms of graph representations and cost functions, to be explained in Section 4.2) to automatically generate referring expressions to refer to each object.

To evaluate the quality of these generated referring expressions, we applied Amazon Mechanical Turk to solicit feedback from the crowd [2]. Through an interface, we displayed an original scene and generated referring expressions (from different generation strategies) in a random order. We asked each turk to select the object in the scene that he/she believed was the one referred to by the shown referring expression (i.e., reference identification task). Each referring expression received three votes from the crowd. In total, 217 turks participated in our experiment.

4.2 Generation Strategies

We applied a set of different strategies to generate referring expressions for each object. The variations lie in two dimensions: (1) different graph representations: using a hypergraph to represent the perceived scene as described in Section 3.1 versus using a regular graph as introduced in (Krahmer et al., 2003); and (2) different cost functions for attributes and relations: cost functions that have been used in previous works (Theune et al., 2007; Krahmer et al., 2008) and cost functions that incorporate uncertainties of perception as described in Section 3.3.2.

Cost functions play an important role in graph-based approaches (Krahmer et al., 2003). Previous works have examined different types of cost functions (Theune et al., 2007; Krahmer et al., 2008; Theune et al., 2011). We adopted some commonly used cost functions from previous work together with the cost functions defined here. In particular, we experimented with the following different cost functions:

Simple Cost: The costs for all hyperarcs are set to 1. With this cost function, the graph-based algorithm resembles the Full Brevity algorithm of Dale (Dale,

[2] To control the quality of crowdsourcing, we recruited participants based on the following criteria: Participants' locations are limited to the United States. Approval rate for each participant's previous work is greater than or equal to 95%, and the number of each participant's previous approved work is greater than or equal to 1000.

1992) in that a shortest distinguishing description is preferred.

Absolute Preferred: The costs for hyperarcs representing absolute attributes (e.g., type, color, and position) are set to 1. The costs for relative attributes (e.g., size) and relations are set to 2. This cost function mimics human's preference for absolute attributes over relative ones (Dale, 1995).

Relative Preferred: The costs for hyperarcs representing absolute attributes are set to 2 and for relative attributes and relations are set to 1. This cost function has been applied previously to emphasize the importance of spatial relations in REG (Viethen and Dale, 2008).

Uncertainty Based: The costs for all hyperarcs are defined by incorporating uncertainties from perception as described in Section 3.3.2.

Uncertainty Relative Preferred: To emphasize the importance of spatial relations as demonstrated in situated interaction (Tenbrink and Moratz, 2003; Kelleher and Kruijff, 2006), the costs for hyperarcs representing relative attributes and relations are divided by 3. This cost function will allow the algorithm to prefer spatial relations through the reduced cost.

Note that we only tested a few (not all) commonly used cost functions proposed by previous work (Krahmer et al., 2003; Theune et al., 2007; Krahmer et al., 2008; Theune et al., 2011). For example, we did not include the stochastic cost function which is defined based on the frequencies of attribute selection from the training data (Krahmer et al., 2003). On the one hand, we did not have a large set of human descriptions of the impoverished scene to learn the stochastic cost. On the other hand, it is not clear whether human strategies of describing the impoverished scene should be used to represent optimal strategies for the robot. Nevertheless, the above different cost functions will allow us to evaluate whether incorporating perceptual uncertainties will make a difference in the REG performance.

4.3 Evaluation Results

As mentioned earlier, each generated referring expression received three independent votes regarding its referent from the crowd. The referent with the most votes is taken as the predicted referent and is used for evaluation. If all three votes are different, then by default, it is deemed that the referent is not correctly identified for that expression. We use the *accuracy* of the referential identification task (i.e., the percentage of generated referring expressions where the referents are correctly identified) as the metric to evaluate different generation strategies illustrated in Section 4.2.

Cost Function	Regular Graph	Hypergraph
Simple Costs	33.2%	33.3%
Absolute Preferred	30.1%	30.3%
Relative Preferred	31.1%	35.4%
Uncertainty Based	35.7%	37.5%
Uncertainty Rel. Prefer.	36.7%	45.2%

Table 1: Results with different cost functions

4.3.1 The Role of Cost Functions

Table 1 shows the results based on different cost functions and different graph representations. There are several observations.

First, when the agent does not have perfect knowledge of the environment and has to automatically infer the environment as in our setting here, cost functions based on uncertainties of perception lead to better results. This occurs for both regular graphs and hypergraphs. This result is not surprising and indicates that cost functions should be tied to the agent's ability to perceive and infer the environment. The uncertainty based cost functions allow the agent to prefer reliable attributes or relations.

Second, consistent with previous work (Viethen and Dale, 2008), we observed the importance of spatial relations. Especially when the perceived world is full of uncertainties, spatial relations tend to be more reliable. In particular, as shown in Table 1, using hypergraphs enables generating group-based relations and results in significantly better performance (45.2%) compared to regular graphs (36.7%) ($p = 0.002$).

Note that our current cost function only includes uncertainties of the agent's own perception in a simplistic form. When humans and agents have mismatched perceptual basis, the human's model of comprehension and tolerance of inaccurate description could play a role in REG. Incorporating human models in the cost function will require in-depth empirical studies and we will leave that to our future

work.

4.3.2 The Role of Imperfect Perception

To further understand the role of hypergraphs in mediating mismatched perceptions between humans and agents, we created a perfect scene regular graph and a perfect scene hypergraph (representing the agent's perfect knowledge of the environment) for each of the 48 scenes used in the experiments. In each of these scene graphs, the attribute and relation descriptors are manually provided. We further applied the *Absolute Preferred* cost function (which has shown competitive performance in previous work) to generate referring expressions for each object. Again, each referring expression received three votes from the crowd.

Table 2 shows the results comparing two conditions: (1) REs generated (by the *Absolute Preferred* cost function) based on the perfect graphs which represent the agent's perfect knowledge and perception of the environment; and (2) REs generated based on automatically created graphs (by the *Uncertainty Relative Preferred* cost function) which represent the agent's imperfect knowledge of the environment as a result of automated recognition and inference. The result shows that given perfect knowledge of the environment, hypergraphs only perform marginally better than the regular graphs ($p = 0.07$). Given imperfect knowledge of the environment, hypergraphs significantly outperforms the regular graphs by taking advantage of spatial grouping information ($p = 0.002$). It is worthwhile to mention that currently we use spatial proximity to identify groups. However, the hypergraph based approach is not restricted to spatial grouping. In theory, it can represent any type of group based on different similarity criteria.

Furthermore, our result shows that the graph-based approaches perform quite competitively under the condition of perfect knowledge and perception. Although evaluated on different data sets, this result is consistent with results from previous work (Gatt and Belz, 2008; Gatt et al., 2009). However, what is more interesting here is that while graph-based approaches perform well when the agent has perfect knowledge of the environment, as its human partner, these approaches literally fall apart with close to 40% performance degradation when applied to the situation where the agent's representation of the shared world is problematic and full of mistakes.

Environment	Regular Graph	Hypergraph
Pefect Perception	80.4%	84.2%
Imperfect Perception	36.7%	45.2%

Table 2: Results of comparing perfect perception and imperfect perception of the shared world.

These results indicate that REG for automatically perceived scenes can be extremely challenging. Many errors result from automated perception and reasoning that will affect the internal representation of the world and thus the generated REs. In our experiments here, we applied a very basic CV algorithm which resulted in rather poor performance in our data: overall, 60.3% of objects in the original scene are mis-recognized, and 10.5% of objects are mis-segmented. We think this poor CV performance represents a more challenging problem.

Some errors such as recognition errors can be bypassed using our current approach based on hypergraphs. For example, in Figure 1 target object 9 (a stapler) and 13 (a key) are mis-recognized as a cup and a pen. Using our hypergraph-based approach, for the target object 9, instead of generating "a small cup" (as in the case of using regular graphs), "a gray object on the top within a cluster of four objects" is generated. For the target object 13, instead of "a pen" as generated by regular graphs, "a small object on the right within a cluster of 4" is generated. Even with recognition errors, these group-based descriptions will allow the listener to identify target objects in their representation correctly. Nevertheless, many processing errors cannot be handled by our current approach. For example, an object can be mistakenly segmented into multiple parts or several objects can be mistakenly grouped into one object. In addition, our current semantic grounding functions are simple. Sometimes they do not provide correct descriptors for the extracted visual features. More sophisticated functions that better reflect human's visual perception (Regier, 1996; Mojsilovic, 2005; Mitchell et al., 2011) should be pursued in the future.

	Minimum Effort	Extra Effort
Pefect Perception	84.2%	88.1%
Imperfect Perception	45.2%	51.5%

Table 3: Results of comparing minimum effort and extra effort using hypergraphs

4.3.3 The Role of Extra Effort

While REG systems have a tendency to produce minimal descriptions, recent psycholinguistic studies have shown that speakers do not necessarily follow the Grice's maxim of quantity, and they tend to provide redundant properties in their descriptions (Jordan and Walker, 2000; Belke and Meyer, 2002; Arts et al., 2011). With this in mind, we conducted a very simple evaluation on the role of extra effort. Once a set of descriptors are selected based on the minimum cost, one additional descriptor (with the least cost among the remaining attributes or relations) is added to the referential description. We once again solicited the crowd feedback to this set of expressions generated by extra effort. Each expression again received three votes from the crowd.

Table 3 shows the results by comparing minimum effort with extra effort when using hypergraphs to generate REs. As indicated here, extra effort (by adding one additional descriptor) leads to more comprehensible REs with 3.9% improvement under perfect perception and 6.3% improvement under imperfect perception (both are significant, $p < 0.05$). The improvement is larger under imperfect perception. This seems to indicate that exploring extra effort in REG could help mediate mismatched perceptions in situated dialogue. However, more understanding on how to engage in such extra effort will be required in the future.

5 Conclusion

In situated dialogue, humans and agents have mismatched perceptions of the shared environment. To facilitate successful referential communication between a human and an agent, the agent needs to take such discrepancies into consideration and generate referential descriptions that can be understood by its human partner. With this in mind, we re-visited the problem of referring expression generation in the context of mismatched perceptions between humans and agents. In particular, we applied and extended the state of the art graph-based approach (Krahmer et al., 2003) in this new setting. Our empirical results have shown that, to address the agent's limited perceptual capability, REG algorithms will need to take into account the uncertainties in perception and reasoning. Group-based information appears more reliable and thus should be modeled by an approach that deals with automated perception of spatially rich scenes.

While graph-based approaches have shown effective for the situation where the agent has complete knowledge of the environment, as its human partner, these approaches are often inadequate when humans and agents have mismatched representations of the shared world. Our empirical results here call for new solutions to address the mismatched perceptual basis. Previous work indicated that referential communication is a collaborative process (Clark and Wilkes-Gibbs, 1986; Heeman and Hirst, 1995). Conversation partners make extra effort to collaborate with each other. For the situation with mismatched perceptual basis, a potential solution thus should go beyond the objective of generating a minimum description, and towards a collaborative model which incorporates immediate feedback from the conversation partner (Edmonds, 1994).

6 Acknowledgments

This work was supported by N00014-11-1-0410 from the Office of Naval Research and IIS-1208390 from the National Science Foundation.

References

Anja Arts, Alfons Maes, Leo Noordman, and Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

E. Belke and A. S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different" decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

H.H. Clark and S.E. Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*, 13:127–149.

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

Madalina Croitoru and Kees Van Deemter. 2007. A conceptual graph approach to the generation of referring expressions. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 2456–2461.

Robert Dale. 1989. Cooking up referring expressions. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, ACL '89, pages 68–75, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Dale. 1992. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* The MIT Press,Cambridge, Massachusetts.

Robert Dale. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263.

Sheel Sanjay Dhande. 2003. A computational model to connect gestalt perception and natural language. In *Masters thesis, Massachusetts Institure of Technology*.

Philip G. Edmonds. 1994. Collaboration on reference to objects that are not mutually known. In *Proceedings of the 15th conference on Computational linguistics - Volume 2*, COLING '94, pages 1118–1122, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. 2004. Generation of relative referring expressions based on perceptual grouping. In *COLING*.

Kotaro Funakoshi, Satoru Watanabe, and Takenobu Tokunaga. 2006. Group-based generation of referring expressions. In *INLG*, pages 73–80.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. 1993. Directed hypergraphs and applications. *Discrete applied mathematics*, 42(2):177–201.

Albert Gatt and Anja Belz. 2008. Attribute selection for referring expression generation: new algorithms and evaluation methods. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 50–58, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albert Gatt, Ielka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The tuna-reg challenge 2009: overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, pages 174–182, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albert Gatt. 2006. Structuring knowledge for reference generation: A clustering algorithm. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, pages 321–328.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 410–419, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42:335–346.

Peter A. Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21:351–382.

Helmut Horacek. 2005. Generating referential descriptions under conditions of uncertainty. In *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG) pages 58-67, Aberdeen, UK*.

Pamela W Jordan and Marilyn Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 181-190.*

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1041–1048, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *computational linguistics*, 38(1):173–218.

Emiel Krahmer Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72, March.

Emiel Krahmer, Mariet Theune, Jette Viethen, and Iris Hendrickx. 2008. Graph: The costs of redundancy in referring expressions. In *In Proceedings of the 5th International Conference on Natural Language Generation, Salt Fork OH, USA*.

Changsong Liu, Rui Fang, and Joyce Y. Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of*

the Special Interest Group on Discourse and Dialogue, SIGDIAL '12, pages 140–149, Stroudsburg, PA, USA. Association for Computational Linguistics.

Changsong Liu, Rui Fang, Lanbo She, and Joyce Y. Chai. 2013. Modeling collaborative referring for situated referential grounding. In *The 14th Annual SIGdial Meeting on Discourse and Dialogue*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2011. Two approaches for generating size modifiers. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of NAACL-HLT 2013, pages 1174-1184*.

Aleksandra Mojsilovic. 2005. A computational model for color naming and describing color composition of images. *IEEE Transactions on Image Processing*, 14:690 – 699.

Terry Regier. 1996. *The human semantic potential*. The MIT Press,Cambridge, Massachusetts.

Deb Roy. 2002. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4.

Alexander Siebert and David Schlangen. 2008. A simple method for resolution of definite reference in a shared visual context. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, SIGdial '08, pages 84–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Sternberg. 2003. *Cognitive Psychology,Third Edition*. Thomson Wadsworth.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. 2012. Referring in installments: a corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 12–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thora Tenbrink and Reinhard Moratz. 2003. Group-based spatial reference in linguistic human-robot interaction. *Spatial Cognition and Computation*, 6:63–106.

Mariët Theune, Pascal Touset, Jette Viethen, and Emiel Krahmer. 2007. Cost-based attribute selection for gre (graph-sc/graph-fp). In *Proceedings of the MT Summit XI Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*.

Mariët Theune, Ruud Koolen, Emiel Krahmer, and Sander Wubben. 2011. Does size matter – how much data is required to train a reg algorithm? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 660–664, Portland, Oregon, USA, June. Association for Computational Linguistics.

Kristinn R. Thrisson. 1994. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 876–881.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 59–67, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Weijers. 2011. Referring expressions with groups as landmarks. volume 15. University of Twente.

Max Wertheimer. 1938. *Laws of organization in perceptual forms. A Source Book of Gestalt Psychology*. Routledge and Kegan Paul, London.

Dengsheng Zhang and Guojun Lu. 2002. An integrated approach to shape based image retrieval. In *Proc. of 5th Asian Conference on Computer Vision (ACCV*, pages 652–657.