Modeling Collaborative Referring for Situated Referential Grounding

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

Abstract

In situated dialogue, because humans and agents have mismatched capabilities of perceiving the shared physical world, referential grounding becomes difficult. Humans and agents will need to make extra efforts by collaborating with each other to mediate a shared perceptual basis and to come to a mutual understanding of intended referents in the environment. In this paper, we have extended our previous graph-matching based approach to explicitly incorporate collaborative referring behaviors into the referential grounding algorithm. In addition, hypergraph-based representations have been used to account for group descriptions that are likely to occur in spatial communications. Our empirical results have shown that incorporating the most prevalent pattern of collaboration with our hypergraph-based approach significantly improves reference resolution in situated dialogue by an absolute gain of over 18%.

1 Introduction

As more and more applications require humans to interact with robots, techniques to support situated dialogue have become increasingly important. In situated dialogue, humans and artificial agents (e.g., robots) are co-present in a shared environment to achieve joint tasks. Their dialogues often involve making references to the environment. To ensure the conversation proceeds smoothly, it is important to establish a mutual understanding of these references, a process called *referential grounding* (Clark and Brennan, 1991): the agent needs to identify what the human refers to in the environment and the human needs to know whether the agent's understanding is correct; and vice versa.

Although reference resolution (Heeman and Hirst, 1995; Gorniak and Roy, 2004; Siebert and Schlangen, 2008) and referential grounding (Traum, 1994; DeVault et al., 2005) have been studied in previous work, the unique characteristics of situated dialogue post bigger challenges to this problem. In situated dialogue, although humans and agents are co-present in a shared world, they have different capabilities in perceiving the environment (a human can perceive and reason about the environment much better than an agent). The shared perceptual basis, which plays an important role in facilitating referential grounding between the human and the agent, thus is missing. Communication between the human and the agent then becomes difficult, and they will need to make extra efforts to jointly mediate a shared basis and reach a mutual understanding (Clark, 1996). The goal of this paper is to investigate what kinds of collaborative efforts may happen under mismatched perceptual capabilities and how such collaborations can be incorporated into our referential grounding algorithm.

Previous psycholinguistic studies have indicated that grounding references is a collaborative process (i.e., *collaborative referring*) (Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991): The process begins with one participant presenting an initial referring expression. The other participant would then either accept it, reject it, or postpone the decision. If a presentation is not accepted, then either one participant or the other needs to refashion it. This new presentation (i.e., the refashioned expression) is then judged again, and the process continues until the current presentation is accepted. To understand the implication of collaborative referring under the situation of mismatched perceptual capabilities, we have conducted experiments on human-human conversation using a novel experimental setup. Our collected data demonstrate an overwhelming use of

Proceedings of the SIGDIAL 2013 Conference, pages 78–86,
Metz, France, 22-24 August 2013.

collaborative referring to mediate a shared perceptual basis.

Motivated by these observations, we have developed an approach that explicitly incorporates collaborative referring into a graph-matching algorithm for referential grounding. As the conversation unfolds, our approach incrementally builds a dialogue graph by keeping track of the contributions (i.e., presentation and acceptance) from both the human and the robot. This dialogue graph is then matched against the perceived environment (i.e., a vision graph representing what are perceived by the robot from the environment) in order to resolve referring expressions from the human. In addition, in contrast to our previous graph-based approach (Liu et al., 2012), the new approach applies hypergraphs: a more general and flexible representation that can capture group-based (n-ary) relations (whereas a regular graph can only model binary relations between two entities). Our empirical results have shown that, incorporating the most prevalent pattern of collaboration (i.e., *agent-present-human-accept*, discussed later) with the hypergraph-based approach significantly improves reference resolution in situated dialogue by an absolute gain of over 18%.

In the following sections, we first give a brief discussion about the related work. We then describe our experiment setting and the patterns of collaboration observed in the collected data. We then illustrate how to build a dialogue graph as the conversation unfolds, followed by the formal definition of the hypergraph representation and the referential grounding procedure. Finally we demonstrate the advantage of using hypergraphs and incorporating a prevalent collaborative behavior into the graph-matching approach for reference resolution.

2 Related Work

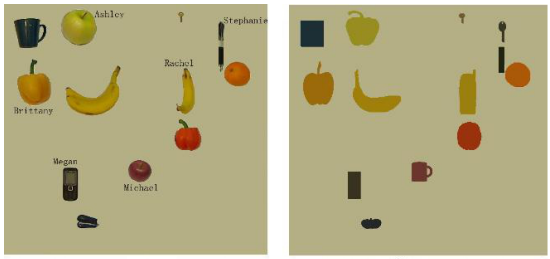
In an early work, Mellish (Mellish, 1985) used a constraint satisfaction approach to identify referents that could be only partially specified. This work illustrated the theoretical idea of how to resolve referring expressions based on an internal model of a world. Heeman and Hirst (Heeman and Hirst, 1995) presented a planning-based approach to cast Clark's collaborative referring idea into a computational model. They used plan construction and plan inference to capture the processes of building referring expressions and identifying their referents. Previous work in situated settings (Dhande, 2003; Gorniak and Roy, 2004; Funakoshi et al., 2005; Siebert and Schlangen, 2008) mainly focused on developing/learning computational models that map words to visual features of objects in the environment. These "visual semantics" of words were then integrated into semantic composition procedures to resolve referring expressions.

These previous work has provided valuable insights in computational approaches for reference resolution. However, they mostly dealt with a single expression or a single referent. In this paper, our goal is to resolve complex referring dialogues that involve multiple objects in a shared environment. In our previous work (Liu et al., 2012), we developed a graph-matching based approach to address this problem. However, the previous approach can not handle group-based relations among multiple objects. Furthermore, it did not look into incorporating collaborative behaviors, which is a particularly important characteristic in situated dialogue. This paper aims to address these limitations.

3 Experiments and Observations

To investigate collaborative referring under mismatched perceptual capabilities, we conducted experiments on human-human interaction (details of the experimental setup can be found in (Liu et al., 2012)). In these experiments, we have two human subjects play a set of naming games. One subject (referred to as the *human-player*) is provided with an original image containing over ten objects (Figure 1(a)). Several of these objects have secret names. The other subject (referred to as the *robot-player*) only has access to an impoverished image of the same scene (Figure 1(b)) to mimic the lower perceptual capability of a robot. The human-player's goal is to communicate the names of target objects to the robot-player so that the robot-player knows which object in his view has what name. The impoverished image was automatically created by applying standard computer vision algorithms and thus may contain different types of processing errors (e.g., mis-segmentation and/or mis-recognition).

Using this setup, we have collected a set of dialogues. The following shows an example dialogue segment (collected using the images shown in Figure 1):



(a) Human-player's view (b) Robot-player's view

Figure 1: An example of different images used in our experiments.

H^1: there is basically a cluster of four objects in the upper left, do you see that
R^2: yes
H: ok, so the one in the corner is a blue cup
R: I see there is a square, but fine, it is blue
H: alright, I will just go with that, so and then right under that is a yellow pepper
R: ok, I see apple but orangish yellow
H: ok, so that yellow pepper is named Brittany
R: uh, the bottom left of those four? Because I do see a yellow pepper in the upper right
H: the upper right of the four of them?
R: yes
H: ok, so that is basically the one to the right of the blue cup
R: yeah
H: that is actually an apple, it is green, I guess it has some amount of yellow on it, but that is a green apple and it is named Ashley
... ...

This example demonstrates two important characteristics regarding referential communication under mismatched perceptual capabilities. First, conversation partners rely on both object-specific properties (e.g., object class, color) and spatial relations to describe objects in the environment. Spatial expressions include not only the binary relations (e.g., "the one to the left of the blue cup"), but also the *group-based* references (Tenbrink and Moratz, 2003; Funakoshi et al., 2005) (e.g., "the upper right of the four of them").

Second, because the shared perceptual basis is missing here, the partners make extra efforts to refer and ground references. For example, the human-player go through step-by-step *installments* (Clark and Wilkes-Gibbs, 1986) to come to the targeted object. The robot-player often proactively provides what he perceives from the environment. The human-player and the robot-player collaborate with each other through iterative *presentation-acceptance* phases as described in the *Contribution Model* proposed in (Clark and Schaefer, 1989; Clark and Brennan, 1991).

[1] H stands for the human-player.
[2] R stands for the robot-player.

These observations indicate that, the approach to referential grounding in situated dialogue should capture not only binary relations but also group-based relations. Furthermore, it should go beyond traditional approaches that purely rely on semantic constraints from single utterances. It should incorporate the step-by-step collaborative dynamics from the discourse as the conversation proceeds.

4 Modeling Collaboration

In this section, we first give a brief description of collaboration patterns observed in our data, and then discuss one prevalent pattern and illustrate how it may be taken into consideration by our computational approach for referential grounding.

4.1 Patterns of Collaboration

Consistent with Clark's Contribution Model, the interactions between the human-player and the robot-player in general fall into two phases: a *presentation* phase and an *acceptance* phase. In our data, a presentation phase mainly consists of the following three forms:

- A complete description: the speaker issues a complete description in a single turn. For example, "there is a red apple on the top right".
- An installment: a description is divided into several parts/installments, each of which needs to be confirmed before continuing to the rest. For example,

  A: under the big green cup we just talked about,
  B: yes
  A: there are two apples,
  B: OK
  A: one is red and one is yellow.

- A trial: a description (either completed or incomplete) with a try marker. For example, "Is there a red apple on the top right?"

In an acceptance phase, the addressee can either accept or reject the current presentation. Two major forms of accepting a presentation are observed in our data:

- Acknowledgement: the addressee explicitly shows his/her understanding, using assertions (e.g., "Yes","Right", "I see") or continuers (e.g., "uh huh", "ok").
- Relevant next turn: the addressee proceeds to the next contribution that is relevant to the current presentation. For example: A says "I see a red apple" and directly following that B says "there is also a green apple to the right of that red one".

In addition, there are also two forms of rejecting a presentation:

- Rejection: the addressee explicitly rejects the current presentation, for example, "I don't see any apple".
- Alternative description: the addressee presents an alternative description. For example, A says "there is a red apple on the top left," and immediately following that B says "I only see a red apple on the right".

In general, referential grounding dialogues in our data emerge as hierarchical structures of recursive presentation-acceptance phases. The acceptance to a previous presentation often represents a new presentation itself, which triggers further acceptance. In particular, our data shows that when mediating their shared perceptual basis, the human-player often takes into consideration what the robot-player sees and uses that to gradually lead to his intended referents. This is demonstrated in the following example[3], where the human-player accepts (Turn 3) the robot-player's presentation (Turn 2) through a *relevant next turn*.

(Turn 1) *H*: There is a kiwi fruit.
(Turn 2) *R*: I don't see any kiwi fruit. I see an apple.
(Turn 3) *H*: Do you see a mug to the right of that apple?
(Turn 4) *R*: Yes.
(Turn 5) *H*: OK, then the kiwi fruit is to the left of that apple.

As shown later in Section 5, this is one prominent collaborative strategy observed in our data. We give this pattern a special name: **agent-present-human-accept** collaboration. Next we continue to use this example to show how the agent-present-human-accept pattern can be incorporated to potentially improve reference resolution.

4.2 An Illustrating Example

In this example, the human and the robot face a shared physical environment. The robot perceives the environment through computer vision (CV) algorithms and generates a graph representation (i.e., a *vision graph*), which captures the perceived objects and their spatial relations[4]. As shown in Figure 2(a), the kiwi is represented as an unknown object in the vision graph due to insufficient object recognition. Besides the vision graph, the robot also maintains a *dialogue graph* that captures the linguistic discourse between the human and the robot.

At Turn 1 in Figure 2(b), the human says "there is a kiwi fruit". Upon receiving this utterance, through semantic processing, a node representing "a kiwi" is generated (i.e., x_1). The dialogue graph at this point only contains this single node. Identifying the referent of the expression "a kiwi fruit" is essentially a process that matches the dialogue graph to the vision graph. Because the vision graph does not have a node representing a kiwi object, no high confidence match is returned at this point. Therefore, the robot responds with a rejection as in Turn 2 (Figure 2(c)) "I don't see any kiwi fruit" [5]. In addition, the robot takes an extra effort to proactively describe what is being confidently perceived (i.e., "I see an apple"). Now an additional node y_1 is added to the dialogue graph to represent the term "an apple" [6]. Note that when the robot generates the term "an apple", it knows precisely which object in the vision graph this term refers to. Therefore, as shown in Figure 2(c), y_1 is mapped to v_2 in the vision graph.

In Turn 3 (Figure 2(d)), through semantic processing on the human's utterance "a mug to the right of that apple", two new nodes (x_2 and x_3) and their relation ($RightOf$) are added to the dialogue graph. In addition, since "that apple"(i.e., x_2) corefers with "an apple" (i.e., y_1) presented by the robot in the previous turn, a coreference link is created from x_2 to y_1. Importantly, in this turn human displays his acceptance of the robot's previous presentation ("an apple") by coreferring to it and building further reference based on it. This is exactly the *agent-present-human-accept* strategy described earlier. Since y_1 maps to object v_2 and x_2 now links to y_1, it becomes equivalent to consider x_2 also maps to v_2. We name a node such as x_2 a **grounded node**, since from the robot's point of view this node has been "grounded" to a perceived object (i.e., a vision graph node) via the agent-present-human-accept pattern.

At this point, the robot matches the updated dialogue graph with the vision graph again and can

[3]This is a clean-up version of the original example to demonstrate the key ideas.

[4]The spatial relations between objects are represented as their relative coordinates in the vision graph.

[5]Note that, since in this paper we are working with a dataset of human-human (i.e., the human-player and the robot-player) dialogues, decisions from the robot-player are assumed known. We leave robot's decision making (i.e., response generation) into our future work.

[6]We use x_i to denote nodes that represent expressions from the human's utterances and y_i to represent nodes from the robot's utterances.

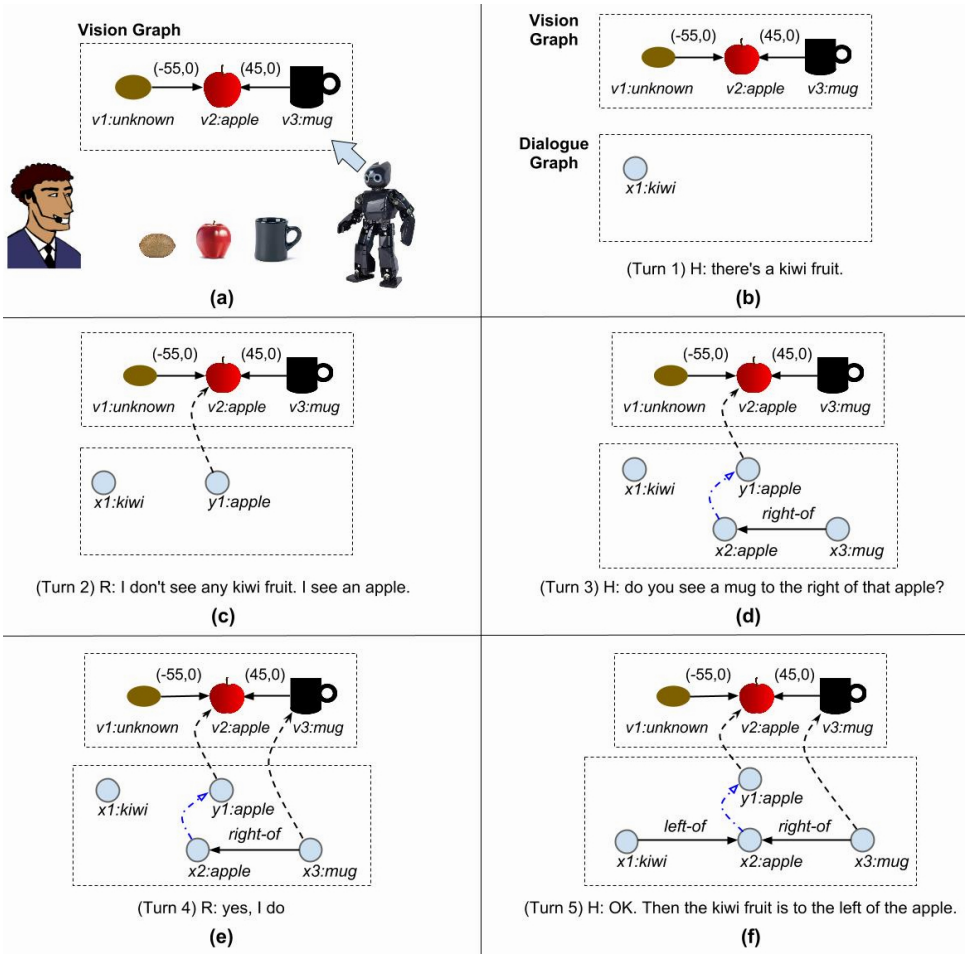


Figure 2: An example of incorporating collaborative efforts in an unfolding dialogue into graph representations.

successfully match x_3 to v_3. Note that, the matching occurs here is considered *constrained graph-matching* in the sense that some nodes in the dialogue graph (i.e., x_2) are already grounded, and the only node needs to be matched against the vision graph is x_3. Different from previous approaches that do not take dialogue dynamics into consideration, the constrained matching utilizes additional constraints from the collaboration patterns in a dialogue and thus can improve both the efficiency and accuracy of the matching algorithm. This is one innovation of our approach here.

Based on such matching result, the robot responds with a confirmation as in Turn 4 Figure 2(e)). The human further elaborates in Turn 5 "the kiwi is to the left of the apple". Again semantic processing and linguistic coreference resolution will allow the robot to update the dialogue graph as shown in Figure 2(f). Given this dialogue graph, based on the context of the larger dialogue graph and through constrained matching, it will be possible to match x_1 to v_1 although the object class of v_1 is unknown.

This example demonstrates how the dialogue graph can be created to incorporate the collaborative referring behaviors as the conversation unfolds and how such accumulated dialogue graph can help referential resolution through constrained matching. Next, we give a detailed account on how to create a dialogue graph and briefly discuss graph-matching for reference resolution.

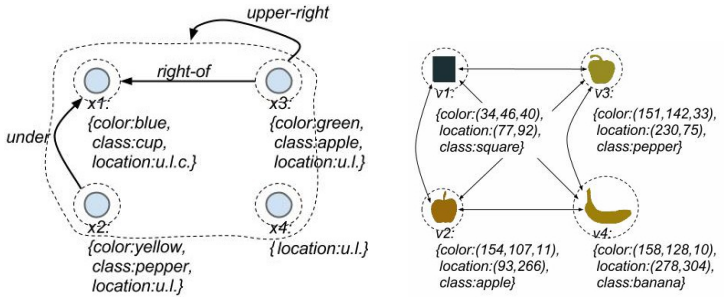
4.3 Dialogue Graph

To account for different types of referring expressions (i.e., object-properties, binary relations and group-based relations), we use hypergraphs to represent dialogue graphs.

4.3.1 Hypergraph Representation

A directed hypergraph (Gallo et al., 1993) is a 2-tuple in the form of $G = (X, A)$, in which

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



(a) Dialogue Graph (b) Vision Graph

Figure 3: Example hypergraph representations

X is a set of nodes, and A is a set of "hyperarcs". Similar to an arc in a regular directed graph, each hyperarc a_i in a hypergraph also has two "ends", i.e., a tail (t_i) and a head (h_i). The tail and head of a hyperarc are both subsets of X, thus they can contain any number of nodes in X. Hypergraph is a more general representation than regular graph. It can represent not only binary relations between two nodes, but also group-based relations among multiple nodes.

For example, suppose the language input issued by the human includes the following utterances:

1. There is a cluster of four objects in the upper left.
2. The one in the corner is a blue cup.
3. Under the blue cup is a yellow pepper.
4. To the right of the blue cup, which is also in the upper right of the four objects, is a green apple.

The corresponding dialogue graph $G_d = (X_d, A_d)$ is shown in Figure 3(a), where $X_d = \{x_1, x_2, x_3, x_4\}$ and $A_d = \{a_1, a_2, a_3\}$. In A_d, for example, $a_1 = (\{x_1\}, \{x_3\})$ represents the relation "right of" between the tail $\{x_3\}$ and the head $\{x_1\}$, and $a_3 = (\{x_3\}, \{x_1, x_2, x_3, x_4\})$ represents the group-based relation "upper right" between one node and a group of nodes.

As also illustrated in Figure 3(a), we can attach a set of labels (or attributes) $\{attr_k\}$ to a node/hyperarc, and use them to store specific information about this node/hyperarc. The perceived visual world can be represented by a hypergraph in a similar way (i.e., a vision graph), as shown in Figure 3(b) [7].

4.3.2 Building Dialogue Graphs

Given the hypergraph representation, a set of operations can be applied to build a dialogue graph as the conversation unfolds. It mainly consists of three components:

Semantic Constraints. Apply a semantic parser to extract information from human utterances. For example, the utterance "*The kiwi is to the left of the apple*" can be parsed into a formal meaning representation as

$$[x1, x2], [Kiwi(x1), Apple(x2), LeftOf(x1, x2)]$$

This representation contains a list of discourse entities introduced by the utterance, and a list of FOL predicates specifying the properties and relations of these entities. For each discourse entity, a node is added to the graph. Unary predicates become the labels for nodes, and binary predicates become arcs in the graph. Group-based relations are incorporated into the graphs as hyperarcs.

Discourse Coreference. For each discourse entity in a referring expression, identify whether it is a new discourse entity or it corefers to a discourse entity mentioned earlier. In our previous example in Figure 2(d), x_2 corefers with y_1, thus a coreference link is added to link the coreferring nodes. Coreferring nodes are merged before matching.

Dialogue Dynamics. Different types of dialogue dynamics can be modeled. In this paper, we only focus on a particularly prevalent type of dynamics as observed from our data, i.e. the agent-present-human-accept pattern as we described in Section 4.1. When such a pattern is identified, the associated nodes (e.g., x_2 in the previous example) will be marked as *grounded nodes* and the mappings to their grounded visual entities (i.e., vision graph nodes) will be added into the dialogue graph.

Based on the above three types of operations, the dialogue graph is updated at each turn of the conversation.

4.3.3 Constrained Matching

Given a dialogue graph $G = (X, A)$ and a vision graph $G' = (X', A')$, reference resolution becomes a graph matching problem which is to

[7] Hyperarcs of the vision graph are not shown in the figure. A hyperarc may exist between any two subsets of objects.

find a one-to-one mapping between the nodes in X and in X'. Due to the insufficiencies of the NLP and the CV components, both the dialogue graph and the vision graph are likely to contain errors. Therefore, we do not require every node in the dialog graph to be mapped to a node in the vision graph, but follow the inexact graph matching criterion (Conte et al., 2004) to find the best match even if they are only partial.

The matching algorithm is similar to the one used in our previous work for regular graphs (Liu et al., 2012), which uses a state-space search approach (Zhang, 1999). The key difference here is to incorporate the agent-present-human-accept collaboration pattern. The search procedure can now start from the state that already represents the known matching of grounded nodes (as illustrated in Section 4.2), instead of starting from the root. Thus it is constrained in a smaller and more promising subspace to improve both efficiency and accuracy.

5 Evaluation

A total of 32 dialogues collected from our experiments (as described in Section 3) are used in the evaluation. For each of these dialogues, we have manually annotated (turn-by-turn) the formal semantics, discourse coreferences and grounded nodes as described in Section 4.3.2. Since the focus of this paper is on incorporating collaboration into graph matching for referential grounding, we use these annotations to build the dialogue graphs in our evaluation. Vision graphs are automatically generated by CV algorithms from the original images used in the experiments. The CV algorithms' object recognition performance is rather low: only 5% of the objects in those images are correctly recognized. Thus reference resolution will need to rely on relations and collaborative strategies.

The 32 dialogue graphs have a total of 384 nodes[8] that are generated from human-players' utterances (12 per dialogue on average), and a total of 307 nodes generated from robot-players' utterances (10 per dialogue on average). Among the 307 robot-player generated nodes, 187 (61%) are initially presented by the robot-player and then coreferred by human-players' following utterances (i.e., relevant next turns). This indicates that the agent-present-human-accept strategy is a prevalent way to collaborate in our experiment. As mentioned earlier, those human-player generated nodes which corefer to nodes initiated by robot-players are marked as grounded nodes. In total, 187 out of the 384 human-player generated nodes are in fact grounded nodes.

To evaluate our approach, we apply the graph-matching algorithm on each pair of dialogue graph and vision graph. The matching results are compared with the annotated ground-truth to calculate the accuracy of our approach in grounding human-players' referring descriptions to visual objects. For each dialogue, we have produced matching results under four different settings: with/without modeling collaborative referring (i.e., the agent-present-human-accept collaboration) and with/without using hypergraphs. When collaborative referring is modeled, the graph-matching algorithm uses the grounded nodes to constrain its search space to match the remaining ungrounded nodes. When collaborative referring is not modeled, all the human-player generated nodes need to be matched.

The results of four different settings (averaged accuracies on the 32 dialogues) are shown in Table 1. Modeling collaborative referring improves the matching accuracies for both regular graphs and hypergraphs. When regular graphs are used, it improves overall matching accuracy by 11.6% ($p = 0.05$, paired Wilcoxon T-test). The improvement is even higher as 18.3% when hypergraphs are used ($p = 0.012$, paired Wilcoxon T-test). The results indicate that proactively describing what the robot sees to the human to facilitate communication is an important collaborative strategy in referential grounding dialogues. Humans can often ground the robot presented object via the agent-present-human-accept strategy and use the grounded object as a reference point to further describe other intended object(s), and our graph-matching approach is able to capture and utilize such collaboration pattern to improve the referential grounding accuracy.

The improvement is more significant when hypergraphs are used. A potential explanation is that those group-based relations captured by hypergraphs always involve multiple (more than 2) objects (nodes). If one node in a group-based relation is grounded, all other involved nodes can have a better chance to be correctly matched.

[8] As mentioned in Section 4.3.2, multiple expressions that are coreferential with each other and describing the same entity are merged into a single node.

	Regular graph	Hypergraph
Not modeling collaborative referring	44.1%	47.9%
Modeling collaborative referring	55.7%	66.2%
Improvement	11.6%	18.3%

Table 1: Averaged matching accuracies under four different settings.

	Group 1	Group 2	Group 3
Number of dialogues	9	11	12
% of grounded nodes	<30%	30%~60%	>60%
Average number of object properties [a]	20	21	12
Average number of relations [b]	11	13	8
Not modeling collaborative referring	49.7%	49.4%	45.3%
Modeling collaborative referring	57.0%	76.6%	63.6%
Improvement	7.3%	27.2%	18.3%

[a] Specified by human-players.

[b] Specified by human-players. The number includes both binary and group-based relations.

Table 2: Matching accuracies of three groups of dialogues (all the matching results here are produced using hypergraphs).

Whereas in regular graphs one grounded node can only improve the chance of one other node, since only one-to-one (binary) relations are captured by regular graphs.

To further investigate the effect of modeling collaborative referring, we divide the 32 dialogues into three groups according to how often the agent-present-human-accept collaboration pattern happens (measured by the percentage of the grounded nodes among all the human-player generated nodes in a dialogue). As shown at the top part of Table 2, the agent-present-human-accept pattern happened less often in the dialogues in group 1 (i.e., less than 30% of human-player generated nodes are grounded nodes). In the dialogues in group 2, robot-players more frequently provided proactive descriptions which led to more grounded nodes. Robot-players were the most proactive in the dialogues in group 3, thus this group contains the highest percentage of grounded nodes. Note that, although the dialogues in group 3 contain more proactive contributions from robot-players, human-players tend to specify less number of properties and relations describing intended objects (as shown in the middle part of Table 2).

The matching accuracies for each of the three groups are shown at the bottom part of Table 2. Since the agent-present-human-accept pattern appears less often in group 1, modeling collaborative referring only improves matching accuracy by 7.3%. The improvements for group 2 and group 3 are more significant compared to group 1. However, group 3's improvement is less than group 2, although the dialogues in group 3 contain more proactive contributions from robot-players. This indicates that in some cases even with modeling collaborative referring, underspecified information from human speakers (human-players in our case) may still be insufficient to identify the intended referents. Therefore, incorporating a broader range of dialogue strategies to elicit adequate information from humans is also important for successful human-robot communication.

6 Conclusion

In situated dialogue, conversation partners make extra collaborative efforts to mediate a shared perceptual basis for referential grounding. It is important to model such collaborations in order to build situated conversational agents. As a first step, we developed an approach for referential grounding that takes a particular type of collaborative referring behavior, i.e. *agent-present-human-accept*, into account. By incorporating this pattern into the graph-matching process, our approach has shown an absolute gain of over 18% in subsequent reference resolution. Extending the results in this paper, our future work will address explicitly modeling the collaborative dynamics with a richer representation. The dialogue graph presented in this paper represents all the mentioned entities and their relations that are currently available at any given dialogue status. But we have not modeled the collaborative dynamics at the illocutionary level. Our next step is to explicitly represent those dynamics, not only for grounding human references to the physical world, but also generating the collaborative behaviors for the agent.

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

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

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

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

Herbert H Clark. 1996. *Using language*, volume 4. Cambridge University Press Cambridge.

Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. 2004. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298.

David DeVault, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 1–4. Association for Computational Linguistics.

Sheel Sanjay Dhande. 2003. A computational model to connect gestalt perception and natural language. Master's thesis, Massachusetts Institute of Technology.

Kotaro Funakoshi, Satoru Watanabe, Takenobu Tokunaga, and Naoko Kuriyama. 2005. Understanding referring expressions involving perceptual grouping. In *4th International Conference on Cyberworlds*, pages 413–420.

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

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

Peter A Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational Linguistics*, 21(3):351–382.

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*, pages 140–149. Association for Computational Linguistics.

Christopher S Mellish. 1985. *Computer interpretation of natural language descriptions*. John Wiley and Sons, New York, NY.

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*, pages 84–87. Association for Computational Linguistics.

Thora Tenbrink and Reinhard Moratz. 2003. Group-based spatial reference in linguistic human-robot interaction. In *Proceedings of EuroCogSci*, volume 3, pages 325–330.

David R Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

Weixiong Zhang. 1999. *State Space Search: Algorithms, Complexity, Extensions, and Applications*. Springer.