Towards Flexible Referring Expression Generation for a Collaborative Robot

Submitted to the faculty of Cognitive Science in partial fulfillment of the requirements for the degree of Bachelor of Science

Kayleigh Bishop

Advised by Brian Scassellati
Yale University
April 2020

Contents

Abstract

1 Introduction **1**

2 Background & Related Work **2**

3 Materials & Methods **3**
3.1 HRC Experimental Setup 3
3.2 Data and Learning Algorithm 4
3.3 Data Collection 5
3.4 Evaluation 6

4 Results **6**
4.1 Collected Data 6
4.2 Model evaluation 7

5 Discussion **8**

6 Conclusions **11**

Acknowledgments

References

Appendix

List of Figures

1 Sample prompt given to participants to gather referring expression data. Participants were given instructions to finish the request for the indicated item. 5
2 Preferences of human participants between human- and model-generated referring expressions. Cases A1 through A3 were grouped together for the purposes of this analysis due to their very similar stimuli and similar distribution of responses. *** indicates significant mean differences; no other case-to-case mean differences were significant. 7
3 Mean relative scoring, by participants who preferred the human-generated string, for model-generated strings. Case C's model string had significantly higher scores on average than any other question ($p < .001$). Error bars indicate 95% confidence intervals. 9
4 Baxter, a robot designed for human-robot collaboration, in the Yale Social Robotics Lab. On the desk is a sample of some task objects from this study that could be recognized by Baxter's object recognition system. 10

List of Tables

1 Data for each prompt from the evaluation survey on Amazon Mechanical Turk. 8

Abstract

A critical aspect of teamwork between humans is the ability to fluently produce both explicit and implicit verbal cues in context. However, speech models for collaborative robots typically rely on simplified speech rules and explicit sentences for communication. In this work, we present a model and training setup for pretrained and online learning of referring expression generation, a task which requires integrating grounded speech with physical and social context. This learning is accomplished by building a dynamic model of reference production from limited training data, which includes information about context in its input. Results indicate that the model can reproduce human-generated referring expressions from training data in most cases, while making errors in particular kinds of tasks that we examine to determine what makes a noun phrase "good" to an English speaker. We also outline potential deployment and assessment of the system on a collaborative teacher-student task between human participants and an autonomous robot.

1 Introduction

The field of Human-Robot Collaboration, or HRC, seeks to design embodied robotic agents that can use their physical abilities to complement the skills of a human worker. The focus of much of the research in the field is to create robots that can collaborate with a human on a joint task typical to manufacturing, such as assembly-line work [10] or furniture assembly [13][18]. While these complex tasks demand communication between partners, collaborative robots are currently very limited in their expressive capacity. State-of-the-art technologies frequently rely on a limited range of predetermined output phrases rather than dynamic language generation [13][17].

The usefulness of fluent communication between collaborators is best exemplified by instances of human-human interaction (HHI). Coordination between human partners is frequently coordinated through natural language. Human partners ground their communication in their shared physical context to easily make requests to one another for support (e.g. "Hold the chair while I attach the leg") or subtask assignment (e.g. "Hand me the flathead screwdriver") [8]. Further, humans specify their language to the particular *social* context of their collaboration via phenomena like lexical entrainment [6] or implicature [1]. This flexible and dynamic nature of human natural language presents a unique challenge to collaborative robots - both in natural language understanding (NLU) and in attempts to replicate it in natural language generation (NLG).

A challenge within the larger framework of NLG is that of referring expression generation (REG), the task of generating noun phrases that identify objects from others in the environment. REG becomes particularly difficult in the application domain of HRC. Collaborative robots are frequently deployed in visually complex and often-changing manufacturing environments. The correct way(s) of referring to a unique object may change as the physical environment changes and new objects are added or removed, and unfamiliar objects may be introduced at any time. This environment makes it impossible to use static referring expressions (REs) for each object - an object that is big in one context may be small in another; an object might be best called a "screwdriver" when by itself, but should be called "yellow flathead screwdriver" when surrounded by others. The ability to use referring expressions is critical to generating natural language that is grounded in the physical and social context of a collaborative task. While HRC environments could be optimized to avoid these challenges by eliminating or simplifying their use of natural language, future generations of robots will be required to collaborate with humans in complex situated interactions, using the communicative means humans are most comfortable with - spoken language [3].

For humans, REG is seemingly effortless on behalf of the speaker. People can generate referring expressions that not only correctly point out the object at hand, but include the precise information, in the precise order, most useful to the listener to help them locate it visually. While a common assumption is that the best RE is the correct one that most efficiently isolates the target [2], evidence suggests the adjective choice and ordering in human REs is rooted in both the particular context of the utterance and humans' perceptual and linguistic capacities [4][20]. A more intuitive example might be one of a task where size and color of tools are both relevant - a speaker might tend to include both size and color in their descriptions, even when only one is needed to pick out the right tool. In contrast, a speaker might hesitate to use the object's pattern if they find the pattern difficult to describe

briefly, and opt to use a different description instead. Additionally, speakers are more likely to overspecify in their descriptions when presented with a larger number of objects [19] or a more distinctive target object [9] The particular structure of REs may reflect the pragmatics of the utterance: English speakers leverage the structure of referring expressions to ease the visual search of the listener [22].

REG serves as an example domain of natural language that requires grounding in the physical and task context in which it is performed. Collaborative robots in complex and unpredictable collaborative settings may need to produce REs that can be easily understood by their human partners, and so need to replicate the perceptually grounded nature of human REs. The application domain of HRC benefits from being somewhat constrained in physical and task context. This could allow a speech-generating robot to learn referring expression behavior unique to multiple (but limited) tasks, environments, or even users.

In this work, we outline a situated REG system that leverages current research in psycholinguistics to learn referring expression behavior from limited training data by integrating speech and contextual information. The proposed system learns to represent the visual search and production qualities of the observed speech with respect to adjective choice and ordering. Novel, human-produced training text is observed from a collaborative task designed to trigger referential speech by requiring particular objects to be isolated from context objects. The implemented language system is evaluated on its ability to reproduce fluent human speech, and ideally would be evaluated in a future experiment on its effectiveness in a collaborative task. Our current experiment to assess its language generation properties demonstrate that the system can dynamically integrate contextual and linguistic information, and adjust its output to suit the unique situation.

2 Background & Related Work

Although HRC systems are not yet capable of emulating the flexibility and fluency of human natural language communication, major leaps have been made in the creation of situated language systems for HRC. Embodied systems have now been implemented that can learn novel words and phrases [18] [6] and spatial terminology [12] [21] from human-robot interaction. Much research in the domain of NLP and NLU focuses on training systems on vast amounts of language data from written corpora, these training examples differ significantly from the two-way and situational natural language interactions that occur in HHI and HRC [16]. In these collaborative contexts, the relevant aspects of objects and the goal of their reference is much more immediate and specific than an abstract reference in written text. Further, systems that rely on large amounts of data are poor fits for embodied robots, for whom data collection is expensive and requires extensive (and tedious) human participation.

One way to increase the amount of data gained from a human interaction is to integrate language with context. A prominent example is multimodal fusion approaches, in which visual and acoustic information are used to better understand human commands [5]. This approach is particularly applicable to the domain of REG, where human natural language is highly dependent on the context of the utterance. Recent data collection of human REs has led to the creation of several REG-centered corpuses. However, the stimuli of these corpuses generally use highly simplified computer graphics images of objects with few distinguishing

features, and prompt descriptions in a non-partnered setting, leading to a distribution of REs that is arguably not reflective of human speech in collaborative situations [16].

Many REG algorithms have been proposed in the past several decades, including Dale and Reiter's Incremental Algorithm and Full Brevity algorithm [2]. Both of these algorithms produce correct specifying REs, but fail to emulate human referring expression production. The Full Brevity algorithm aims to avoid redundant descriptions and find the shortest possible description - a task that is computationally expensive, and not an accurate depiction of human linguistic behavior. The more relaxed Incremental Algorithm assumes a fixed preference order over descriptive features, adding them in order if they have any descriptive value. While this algorithm can sometimes provide redundant descriptions as humans do, the assumption of a fixed preference order subverts any understanding of the human partner's perspective, as well as task- and setting-relevant information [14].

We present a system that leverages situated machine learning to acquire data about human referring expression behavior. The benefit of this approach is that it provides access to both speech data and contextual information. It also allows the agent to probe RE behavior in its specific collaborative setting, rather than solely learning from monologue human-produced descriptions. The system relies on studies of human referring expression generation, which indicate that adjective choice is influenced by both the discriminative quality of the feature and the perceptual qualia of the speaker [20].

3 Materials & Methods

3.1 HRC Experimental Setup

The system presented here was designed with the intent to be deployed on a robot in an interactive collaborative task with a human partner. The data collection and training focused on REG behavior with respect to different contextual arrangements of basic objects.The task objects are designed to be easily recognizable and with easily labeled features in the dimensions of color, size, and length. These objects include screwdrivers, colored blocks, water bottles, pens, and cans. The objects are brightly painted but otherwise lack specifying features besides their shape and size.

This experiment also required that the system assign appropriate adjectives to object color, size, and shape. For color, we used a straightforward implementation of estimated rgb-to-label mapping from [7]. Size and shape present a particular challenge as being object type- and context-dependent. For the purposes of this experiment, we chose to make all stimuli relatively easily contrasted in shape and size; the speech model's confidence was then based on the normalized difference in dimensions between objects of the same type. This choice was based on semantic characterizations of gradable descriptors about size and length [11], but could and should be updated to capture more nuance (see Discussion). The final implementation and trained model is available at **https://github.com/kayleigh-bishop/hrc_source.**

3.2 Data and Learning Algorithm

We propose an approach for referring expression selection which incrementally selects adjectives at each stage. This selection process is enabled by maintaining distinct sub-models for each object feature. These feature models take as input both the distribution of the feature and its assigned label in the given environment and the speech model's belief that the particular label it chose for that feature is correct. This feature distribution parameter is determined by the discrimination factor of the particular feature - i.e., what proportion of objects in the given context it would eliminate. This choice of parameter was based on experimental evidence of human RE production [20] and encouraged by our preliminary assessment of training data. The overall speech model then incorporates these separate feature models to provide judgements on the best term to apply. For a given object and context, the model then outputs the best term to use next (or the empty string if none is required. Put more formally, for an observed context $c_t \in C$ and target object $o \in O$, the adjective term $\hat{a}_f$ chosen by the system at time step t is the term that maximizes:

$$\hat{a}_f = \arg\max_{a_f \in f(o,c_t)} [p_f(a_f|o,c_t) * p_s(a_f|o,c_t)] \tag{1}$$

where $f(o,c)$ represents the possible feature labels that could be assigned to object o in context c; p_f represents the feature model, which determines if terms are included in output; and p_s represents the speech model's confidence that the label assignment a_f is accurate.

The speech model was required to assign appropriate adjectives and confidence levels to object color, size, and shape. For color, we used a straightforward implementation of estimated RGB-to-label mapping from [7]. Size and shape present a particular challenge as being object type- and context-dependent. For the purposes of this experiment, we chose to make all stimuli relatively easily contrasted in shape and size; the speech model's confidence was then based on the normalized difference in dimensions between objects. Although the state of all possible object descriptions is intractable, we chose to limit the possibilities to color, size, and length characteristics, due to their commonality in REG datasets as well as their relative ease of perceiving via computer vision. Over the entire training image set, all of the 11 basic colors were included as object features; size and length varied on a continuous basis per object and stimulus, but generally fell into 2-3 distinct categories when used contrastively.

Here, we take *context* or $c_t \in C$ to mean the set of objects that are i) observable by the system and/or robot and ii) have *not* already been eliminated by a term earlier in the speech string at time step t. This incremental approach is supported by psychophysical accounts of human overspecification in reference production [14]. While the space of all possible object sets C is also intractable, we model each context by extracting the normalized discrimination factor from it, as a representation of the feature distribution in the context. This allows the model to make judgements about totally unfamiliar contexts by relating the feature distribution to those it has seen before.

We used a logistic regression model to represent $Pr(a_f|o,c_t)$. The context parameter is converted to a simple 2D vector, containing the discrimination term d_f and the total number of objects in consideration, $N_c = |c_t|$. For a given resulting vector x and the logistic regression model's parameter vector θ_f, the estimate of the term usage probability is given

by:

$$p_f(a_f|o, c_t) = \frac{C}{1 + e^{-\theta_f^T x}} \tag{2}$$

where C is a normalization constant.

3.3 Data Collection

Figure 1: Sample prompt given to participants to gather referring expression data. Participants were given instructions to finish the request for the indicated item.

Baseline and training data was collected from a survey on Amazon's Mechanical Turk platform over two phases. Two surveys were used - the first containing 25 stimuli, and the second containing 10. Different participants were recruited for each survey. All participants were required to speak English fluently. Each survey contained digitally edited stimuli composed of edited photographs of the stimuli objects to create novel scenes on an apparent 2D plane. Survey 1 contained a brief evaluative section that was administered before the rest of the task to ensure the participant understood the instructions; otherwise, stimuli were shown to the participant in a randomized order, with one stimuli and text box per page. In each stimulus image, one "target object" was indicated by a large white arrow.

Participants first received instructions for the task, which indicated that a partner would be using their descriptions to pick the correct item in another task. Instructions were formulated to not include any actual descriptions that could apply to objects in the task, to avoid priming or biasing participants to respond in a particular way. After receiving instructions, the participants were presented with the task: one stimulus image per page, with a text box prompt asking them to finish the sentence "Pass me the...[blank]." The goal was to induce a collaborative dialogue-like task as much as possible while collecting data online.

Importantly, free-response descriptions were collected and used in an unedited form. Responses were included as long as the participant involved demonstrated understanding of the task by providing unambiguous descriptions to at least two-thirds of the stimuli. In total, 1,736 usable stimulus-label pairs were collected from 92 participants, with responses from 33 participants being excluded for failing to meet the criteria above.

3.4 Evaluation

The original plan to evaluate this system involved an experimental setup with a Baxter robot and an in-person human participant, in which the system would be used to communicate with the human partner about objects needed for the collaborative experimental task. However, due to the COVID-19 pandemic that resulted in laboratory closures, in-person experiments involving laboratory robots could not be conducted as planned.

To establish some assessment of the fluency and acceptability of the REs produced by the system, an Amazon Mechanical Turk study similar to those used for data collection was designed. This study was designed to compare the modal human-generated REs with those generated by the model. The model was given the same data used in training for each stimulus, which included info on object identities, rgb values, and bounding boxes. Importantly, the stimuli used in this study were exclusively those for which the human and model responses differed.

In the survey, participants were presented an image stimulus from the previous data collection studies. The participants were then asked to choose the phrase that they thought best fit the stimulus. After this choice they were asked to score their second choice (the non-preferred description) out of 100 points, with 100 points meaning the second choice was equally as good as their preferred choice. The survey consisted of 6 such casesin total: 3 representing shape vs. size contrasts in descriptions (the most common error in the system's REG) (A1, A2, A3; 1 representing adjective ordering differences (B); 1 representing redundancy vs. brevity (C); and 1 representing size vs. length usage (D).

4 Results

4.1 Collected Data

In this experiment we collected a variety of user free-response descriptions of a limited set of objects in various digitally-composed contexts. Because psychophysics research has suggested that discriminability partially determines adjective usage and ordering [20], and our model relies on this assumption, we ran a series of linear regression tests using the

proportional discriminability scores of color, size, and length as predictors to the usage of all three feature types in responses.

There was a significant positive association between the proportional discriminability of color and the use of color in responses, $B = .424, t(33) = 2.846, p = .008$, holding size and length distributions constant. There was also a significant positive association between the proportional length discriminability and the usage of height/length in responses, $B = 24.60, t(33) = 3.819, p = .001$, holding color and size distributions constant. There was no significant association between proportional size discriminability and size usage, $B = .278, t(33) = 1.184, p = .245$, when controlling for color and length distributions. Further, there were no significant associations between distributions of one feature and usage of another when accounting for Bonferroni correction (all associations $p >= .030$). We elected not to analyze the role of object count in usage, both because our experimental setup contained minimal controlled variation in object count and because the influence of object count on object specification has been well-established elsewhere [22][19][9].

4.2 Model evaluation

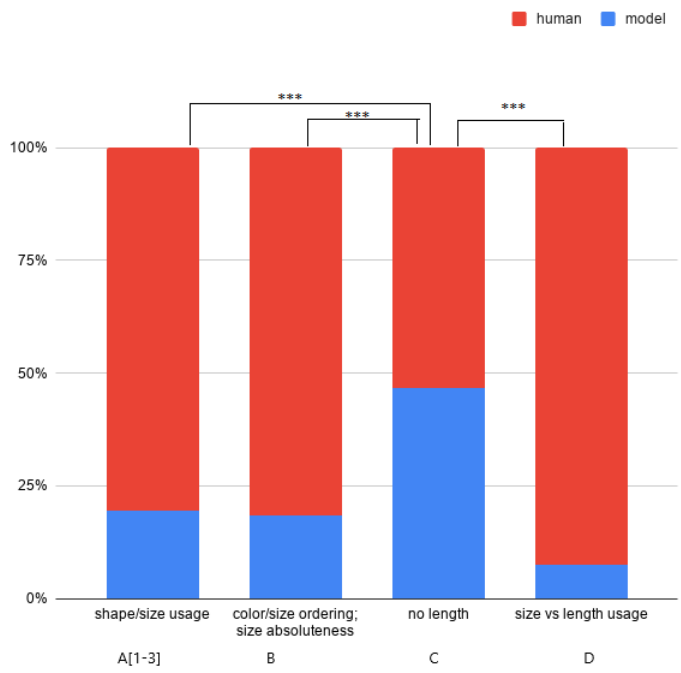

Figure 2: Preferences of human participants between human- and model-generated referring expressions. Cases A1 through A3 were grouped together for the purposes of this analysis due to their very similar stimuli and similar distribution of responses. *** indicates significant mean differences; no other case-to-case mean differences were significant.

In order to get a general sense of the model's accuracy after training, we compared the referring expressions it generates in simulation to the descriptions given by users in the training phase of the evaluation. The model reproduced the exact adjective use and ordering of the most common human-generated descriptions for 14 of the 20 training stimuli. The 6

Prompt	Description	Human-string	Model-string	Preferred human string (%)	Model mean relative score	SD
A1	Cube distinguished by size/shape and color	purple cube block	purple small block	79.4	57.8	25.3
A2	Long block specified by size/shape and color	purple rectangular block	purple big block	82.9	54.2	28.6
A3	Cube distinguished by size/shape and color	purple cube block	purple small block	79.3	56.7	27.2
B	Cube distinguished by size and color	smallest green block	green small block	81.7	53.1	28.8
C	Long block specified by size/shape or color	purple long block	purple block	53.4	66.4	24.4
D	Long block specified poorly by color, and well by size/shape	purple long block	purple big block	92.4	54	27.3

Table 1: Data for each prompt from the evaluation survey on Amazon Mechanical Turk.

stimuli on which it differed were incorporated into the comparison study conducted on Amazon Mechanical Turk to gauge the severity of the discrepancy in the produced descriptions. Overall, in these disputing cases there was a strong preference towards the human-generated strings, which were chosen 78.2% of the time ($N = 1509$).

A one-way analysis of variance yielded a main effect for the question type, $F(5, 1503) = 27.4, p < .001$. A post-hoc Tukey's HSD analysis showed preference towards human-generated strings was significantly lower in the redundant adjective case (case C) than in all other cases, $p < .001$, dropping down to near-chance levels of 53.4%. There were no significant differences in preference tendencies for any other stimuli.

Among those who preferred the human-generated descriptions, overall scores for the model descriptions varied widely between participants ($M = 58.2, SD = 27.2$). A one-way analysis of variance on these scores again indicated a main effect of question type, $F(5, 1503) = 5.00, p < .001$. Post-hoc HSD comparisons indicated that case C had a significantly higher scoring model string ($M = 67.1, SD = 26.1$), $p < .001$, than all cases other than A1. There were no significant differences in average model string scores between any other cases.

5 Discussion

In this study we collect and assess referring expression training data and perform some preliminary evaluation of a new referring expression generation system. Analysis of trends in the collected RE data provide some initial confirmation that there is a linear relationship between discriminability of a feature and its usage, as predicted by background work, which could be leveraged in a computational approach such as ours.

The model was able to learn to replicate human REs relatively consistently on stimuli which varied in the values and frequency of given features and in the (subjective) value of redundant information. This included choosing generally correct color and size adjectives using the relatively simplistic semantic system provided to it in its implementation, as well as choosing correct ordering and amount of adjectives.

That said, it proved valuable to examine more closely the cases in which the model failed to accurately replicate human descriptions, and the kinds of errors that participants

generally found tolerable or intolerable. Case C, which contrasted a model's brief noun-phrase with a human one that provided additional redundant information, performed the best, with preference and scoring results that suggested people overall found the descriptions more equal in quality - perhaps varying solely on personal preference of the participant. In contrast, other cases had consistently low scores among both preference and scoring results, indicating that the model's output was not acceptable compared to human speech.

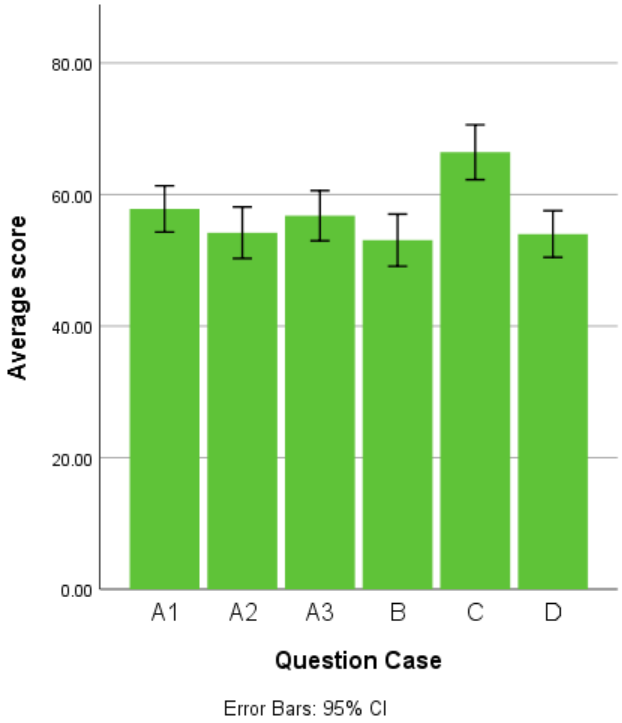

Figure 3: Mean relative scoring, by participants who preferred the human-generated string, for model-generated strings. Case C's model string had significantly higher scores on average than any other question ($p < .001$). Error bars indicate 95% confidence intervals.

The set of cases A1-A3 highlighted the model's inability to use fixed shape adjectives in its descriptions, such as "cube" or "rectangular", which humans used frequently; instead, the model compensated using size adjectives to discriminate the correct object. An additional issue was that the object's overall size (either "large" or "small") was not always the most accurate way of describing its dimensions, compared to an adjective of its length or shape. For example, a rectangular block may be more massive and technically larger than a cube of its same width and height, but the results here suggest that using "larger" is much less acceptable than referring more specifically to its length or shape. Case D highlights this distinction more clearly, contrasting a human's description of "long" with the model's description of "large" for the same target object as in A1; participants generally found the model's description less acceptable, giving it a mean score of 54 out of 100. In general, future versions of this model could be improved by adding more kinds of adjectives, including shape; one might also want to include more instances of "long" vs. "short" objects, and other such dimensional contrasts, in the training data.

Case B contained two contrasts of interest, highlighting the model's inability to include

superlative adjectives ("smallest" vs. "small") as well as a distinction in ordering of color and size adjectives (human string "smallest green" vs. model string "green small"). While English adjective ordering is flexible and context-dependent, as shown in [20], there seems to be some intuitive default, so to speak, of ordering, which "green small" appears to violate. In theory, this ordering trend should be captured in the training data, but the inclusion of the superlative adjective further complicates the issue. Superlatives like "smallest" are by definition more discriminating than their simple forms. This likely has an effect on both ordering phenomena and preference tendencies among participants; this case highlights the need for future versions to use and account for superlatives in training and deployment.

A strength of this system is its ability to extract a relatively large amount of data from a small pool of examples. For example, from the description "The small red block," the system could extract three separate data points for each feature model (9 in total). While the inclusion of more features than the three used in this experiment would come at a computational cost, the amount of comparative data points that could be extracted from a single utterance would increase by one for each feature model. This makes adding more features to speech consideration, without scaling up data collection significantly, a viable option for future work. Further, having the model extract generalized training data (*should*) from experience allows it to generalize well to completely unfamiliar contexts, as well as make quick adjustments between users and environments.

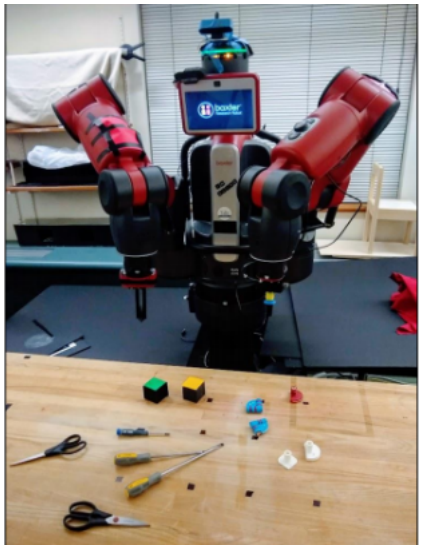

Figure 4: Baxter, a robot designed for human-robot collaboration, in the Yale Social Robotics Lab. On the desk is a sample of some task objects from this study that could be recognized by Baxter's object recognition system.

It would be best to evaluate this system in an embodied, collaborative environment as was planned. This would be preferable because human speech behavior varies based on the context and goal of the conversation, as discussed. Thus the effectiveness of an HRC-focused speech generation system could only be evaluated in a human-robot collaborative context - even a simple one, such as an object sorting task. A weakness of the deployed system as it was planned would potentially be a simplistic perceptual system. Pre-trained 2D object recognition were used to recognize objects, while the resulting bounding boxes were used to

extract data on color and size. Thus, the failure of object recognition on any object in a workspace leads to the model working off incorrect information and, in the worst case, failing to identify the target object at all. A system deployed in an environment with a constrained range of types of objects (though not necessarily features of those objects) could use, for example, a specifically trained deep neural network to achieve more reliable results.

Another opportunity allowed by deployment on a collaborative robot could be to employ brief training sessions with each partner and/or for each task. It could be interesting to assess the value of "tuning" the model after corpus training, to allow it to adapt more dynamically to different task situations or user preferences.

Future work could seek to include include the integration of implementations of lexical entrainment[6]. A robot with on-line learning of both semantic grounding and description production could better participate in the social micro-instances of language Wittgenstein dubbed "language games." Similarly, future work could integrate spatial reference, which includes perspective taking [15]. In general, recent work at the intersection of HRC and natural language processing has, and should continue to, more closely examine the relationship between language and cognition in human collaborative settings. This work presents a small example of how building a dynamic mental model that integrates language and context can improve progress towards creating more human-like, communicative intelligent systems.

6 Conclusions

In this paper we present a flexible, psychophysically-inspired system capable of building incremental models of referring expressions by integrating speech/text and context during training. After training on human-produced data, the model successfully replicated the most common human-generated description for 14 out of 20 training stimuli, which included a variety of distractor objects. In experimentally assessing the model on stimuli in which it differed from the most common human response, we found that varying levels of redundancy is well tolerated, but more nuance of length, size, shape, and superlative adjectives is needed for reliable performance.

Future work will focus on extended this system as mentioned above and deploying it on a robot in a collaborative context. It may also include additional mini-training sessions with each user and/or task to see if the extra, more locally specific data results in significant changes in generated noun phrases.

Acknowledgments

I am incredibly grateful to Professor Brian Scassellati for his guidance, support, and advice throughout this research process, and his unfailing tolerance of my struggling with basic research tasks; to Jake Brawer for his patience and his help in making this project possible; to all the amazing members of Scazlab that got me involved in robotics research; to the Yale Computer Science Department for supporting my research and instilling the fear of God in me; and to all the members of the Yale Cognitive Science program for providing a major and thesis process I actually liked and (presumably) letting me graduate.

References

[1] H. Grice, "Logic and Conversation", pp. 85–113, 1968, ISSN: 00218901. DOI: 10.1057/9780230005853_5. arXiv: arXiv:1011.1669v3. [Online]. Available: http://link.springer.com/10.1057/9780230005853%7B%5C_%7D5.

[2] R. Dale and E. Reiter, "Computational interpretations of the Gricean maxims in the generation of referring expressions", *Cognitive Science*, vol. 18, pp. 233–263, 1995, ISSN: 03640213. DOI: 10.1016/0364-0213(95)90018-7.

[3] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration", *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004*, vol. 3, pp. 1030–1037, 2004.

[4] E. Belke, *Visual determinants of preferred adjective order*, 3. 2006, vol. 14, pp. 261–294, ISBN: 1350628050. DOI: 10.1080/13506280500260484.

[5] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski, "Using vision, acoustics, and natural language for disambiguation", *HRI 2007 - Proceedings of the 2007 ACM/IEEE Conference on Human-Robot Interaction - Robot as Team Member*, pp. 73–80, 2007. DOI: 10.1145/1228716.1228727.

[6] T. Iio, M. Shiomi, K. Shinozawa, T. Miyashita, T. Akimoto, and N. Hagita, "Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary?", *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, pp. 3727–3734, 2009. DOI: 10.1109/IROS.2009.5354149.

[7] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications", *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1512–1523, 2009, ISSN: 10577149. DOI: 10.1109/TIP.2009.2019809.

[8] J. Shah and C. Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human-robot teaming", *Human Factors*, vol. 52, no. 2, pp. 234–245, 2010, ISSN: 00187208. DOI: 10.1177/0018720809350882.

[9] R. Koolen, A. Gatt, M. Goudbeek, and E. Krahmer, "Factors causing overspecification in definite descriptions", *Journal of Pragmatics*, vol. 43, pp. 3231–3250, 2011, ISSN: 03782166. DOI: 10.1016/j.pragma.2011.06.008.

[10] R. A. Knepper, T. Layton, J. Romanishin, and D. Rus, "IkeaBot: An autonomous multi-robot coordinated furniture assembly system", *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 855–862, 2013, ISSN: 10504729. DOI: 10.1109/ICRA.2013.6630673.

[11] C. Qing and M. Franke, "Gradable adjectives , vagueness , and optimal language use :" *Proceedings of SALT*, vol. 24, pp. 23–41, 2014.

[12] J. Tan, Z. Ju, and H. Liu, "Grounding spatial relations in natural language by fuzzy representation for human-robot interaction", *IEEE International Conference on Fuzzy Systems*, pp. 1743–1750, 2014, ISSN: 10987584. DOI: 10.1109/FUZZ-IEEE.2014.6891797.

[13] B. Hayes and B. Scassellati, "Effective robot teammate behaviors for supporting sequential manipulation tasks", *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-Decem, pp. 6374–6380, 2015, ISSN: 21530866. DOI: 10.1109/IROS.2015.7354288.

[14] R. Koolen, E. Krahmer, and M. Swerts, "How Distractor Objects Trigger Referential Overspecification: Testing the Effects of Visual Clutter and Distractor Distance", *Cognitive Science*, vol. 40, no. 7, pp. 1617–1647, 2016, ISSN: 15516709. DOI: 10.1111/cogs.12297.

[15] S. Li, R. Scalise, H. Admoni, S. Rosenthal, and S. S. Srinivasa, "Spatial references and perspective in natural language instructions for collaborative manipulation", *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 44–51, 2016. DOI: 10.1109/ROMAN.2016.7745089.

[16] D. d. S. Rocha and I. Paraboni, "Reference production in human-computer interaction : Issues for Corpus-based Referring Expression Generation", pp. 2994–2998, 2016.

[17] A. Roncone, O. Mangin, and B. Scassellati, "Transparent role assignment and task allocation in human robot collaboration", *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1014–1021, 2017, ISSN: 10504729. DOI: 10.1109/ICRA.2017.7989122.

[18] J. Brawer, O. Mangin, A. Roncone, S. Widder, and B. Scassellati, "Situated Human-Robot Collaboration: Predicting intent from grounded natural language", *IEEE International Conference on Intelligent Robots and Systems*, pp. 827–833, 2018, ISSN: 21530866. DOI: 10.1109/IROS.2018.8593942.

[19] M. Elsner, A. Clarke, and H. Rohde, "Visual Complexity and Its Effects on Referring Expression Generation", *Cognitive Science*, vol. 42, pp. 940–973, 2018, ISSN: 15516709. DOI: 10.1111/cogs.12507.

[20] K. Fukumura, "Ordering adjectives in referential communication", *Journal of Memory and Language*, vol. 101, pp. 37–50, 2018, ISSN: 0749596X. DOI: 10.1016/j.jml.2018.03.003.

[21] C. D. Wallbridge, S. Lemaignan, E. Senft, and T. Belpaeme, "Towards Generating Spatial Referring Expressions in a Social Robot: Dynamic vs. Non-Ambiguous", *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 616–617, 2019. arXiv: 1806.03831. [Online]. Available: http://arxiv.org/abs/1806.03831.

[22] P. Rubio-Fernandez, F. Mollica, and J. Jara-Ettinger, "English and Spanish speakers exploit language structure to increase communicative efficiency", *Preprint*, 2020.

Appendix

Text instructions for data collection

Welcome! In this study, we are interested in how you describe common objects in different situations. Please read the instructions below carefully.

You will be presented with a series of 10 pictures. The pictures will contain common household items in a workspace. One object will be indicated by a large white arrow.

Your task is to give instructions to a partner that will be looking at the workspace pictured, so that they can pick out the correct specific item indicated by the arrow.

In the text box below each picture, complete the sentence by describing the object to your partner.

The objects you will see in this task include wooden blocks, screwdrivers, scissors, mugs, water bottles, and pens.

Please answer as quickly and naturally as you can. The task is brief, so we expect you will not be rushed.

The task will begin on the next page.

Text instructions for online evaluation

Pre-experiment introduction

The estimated completion time for this survey is 3 minutes.

In each question, you will look at a picture of several household objects. One object - the target object - is indicated by a large white arrow.

You will then read two different requests to get the target object. Please assess and grade the two descriptions.

Per-question instructions (sample)