Generating Quantified Referring Expressions through Attention-Driven Incremental Perception

Gordon Briggs
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
Washington, DC USA 20375
gordon.briggs@nrl.navy.mil

Abstract

We model the production of quantified referring expressions (QREs) that identity collections of visual items. A previous approach, called Perceptual Cost Pruning, modeled human QRE production using a preference-based referring expression generation algorithm, first removing facts from the input knowledge base based on a model of perceptual cost. In this paper, we present an alternative model that incrementally constructs a symbolic knowledge base through simulating human visual attention/perception from raw images. We demonstrate that this model produces the same output as Perceptual Cost Pruning. We argue that this is a more extensible approach and a step toward developing a wider range of process-level models of human visual description.

1 Introduction

Modeling the generation of human-like referring expressions in visual contexts is an ongoing challenge in the field of natural language generation (NLG). One key aspect of this challenge is the disparity between how humans and current computational approaches operate at a process-level. Humans have perceptual and cognitive limitations; they can not mentally represent visual scenes down to the exact detail of every visual object or collection of objects. Thus, people tend to selectively attend to visual scenes in order to acquire enough information to complete their task (Yarbus, 1967).

In contrast, current referring expression generation (REG) algorithms generally assume a fully-specified, symbolic knowledge base (Van Deemter, 2016). Essentially, these approaches abstract away the process of perception and seek to model patterns of human language primarily at the content selection phase. One task where this approach breaks down is quantified reference expression (QRE) generation.

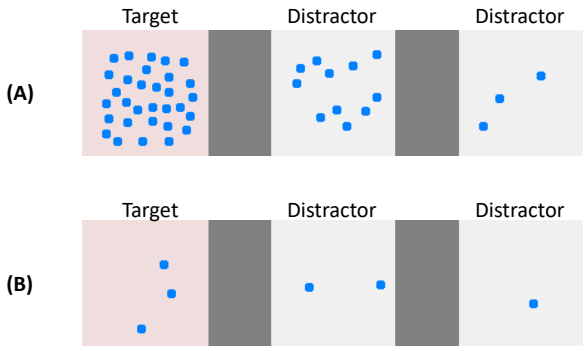

Figure 1: Examples of QRE generation task problems, based on examples from (Barr et al., 2013).

QRE tasks involve referring to collections of visual items by communicating information about the quantities contained in each collection. Initial experimental work in human-produced QREs showed regularities in responses that were not easily explained by content selection processes (Barr et al., 2013). To illustrate this, consider the two examples of QRE problems found in Figure 1. In Problem A, it was found that participants favored *relative* quantity expressions (e.g., "the box with the *most* circles") over *exact* number expressions (e.g., "the box with *thirty-one* circles"). In contrast, participants favored exact expressions in Problem B (Barr et al., 2013).

To account for this finding, the experimenters appealed to a principle of least perceptual effort. Determining the exact number of approximately 30 objects is effortful and time-consuming, whereas determining that there are exactly three objects is quick and requires little effort. Thus, the findings can be explained by people generating exact QREs so long as determining the exact quantity did not exceed some threshold of effort.

This explanation was validated using a computational method called Perceptual Cost Pruning, proposed previously by (Briggs and Harner, 2019).

Proceedings of The 13th International Conference on Natural Language Generation, pages 107–112,
Dublin, Ireland, 15-18 December, 2020.

Perceptual Cost Pruning was able to model human QRE production by: (1) starting with a complete symbolic knowledge base representing the visual scene; (2) removing facts from the input knowledge base based on a model of the time cost of exact enumeration; (3) using a preference-based referring expression generation algorithm (i.e., the Incremental Algorithm) on this reduced knowledge base. We call this approach a *destructive* approach, wherein a full knowledge base is degraded to be sparser.

In this paper, we present a *constructive* approach to modeling the production of QREs, wherein a sparse, symbolic scene representation is built up from nothing. Specifically, we present an alternative model that incrementally builds a symbolic knowledge base through simulating human visual attention/perception from raw images. We demonstrate that this model produces the same output as perceptual cost pruning. Furthermore, we argue that this is a more extensible approach and a step toward developing a wider range of process-level models of human visual description.

2 The QRE Task

Stimuli To model the QRE task we use the stimuli described by (Barr et al., 2013). These stimuli constitute 20 QRE generation problems, each with one target collection and two distractor collections. The items in each collection are all the same with respect to size, shape, color, and all other attributes. They are also randomly distributed in their respective containers. Thus, the total quantity of items in each collection becomes a salient feature to differentiate collections, instead of individual object attributes or collection attributes like structured arrangement (e.g., shape of the group of items). While the precise quantities for each target and distractor pair were available, the original images were not. As such, we constructed a series of 120 images, six for each problem, corresponding to the different possible target and distractor configurations. Examples of these images can be found in Figure 1.

Output In the original study (Barr et al., 2013), participants' QREs were annotated as following into different categories, and these annotations were refined in subsequent work (Briggs and Harner, 2019). We use the latest annotation categories, which are: Exact Number (`NUM`) (e.g., "the box with 31 dots"); Relative quantity (`REL`) (e.g., "the box with the most dots"); and Absolute description (`ABS`) (e.g., "the box with dots"). The QRE generation algorithm presented in this paper is designed to predict which of these categories is included given a particular QRE task image.

3 A Constructive Model of QRE Generation

Our model of QRE generation was developed within the ARCADIA cognitive framework (Bridewell and Bello, 2016). ARCADIA provides an ideal framework upon which to implement a model of incremental perception and quantified description for the following reasons: (1) attention and its strategic control is the central organizing feature of the system; (2) it provides the representational flexibility necessary for modeling both an approximate number system that supports group quantity estimation and an object-tracking system that supports exact enumeration (i.e., counting); (3) it aims to implement a cognitively plausible model of human visual processing; and (4) it operates in discrete cycles, allowing for modeling the time course of perceptual and cognitive processes. In the interest of space, we omit more precise technical details regarding ARCADIA models, which can be found in other work (Bridewell and Bello, 2016; Lovett et al., 2019).

3.1 Representations

To produce the range of exact, relative, or absolute quantified expressions requires different forms of numerical representation at different levels of precision. Research shows that multiple forms of numerical representation underlie human number sense (Feigenson et al., 2004). Approximate representations are obtained quickly through estimation (Barth et al., 2003), while exact representations are obtained slowly through counting for quantities outside the subitizing range of about four items (Gelman and Gallistel, 1986). Furthermore, evidence suggests that approximate and exact representations of quantity are obtained through different ways of deploying spatial attention (Hyde and Wood, 2011). Attention to groups results in estimation and approximate representation of quantity, while attention to individual objects underlies exact enumeration.

In ARCADIA, approximate quantities are represented as Gaussians with a mean, $\mu = n$, equal to

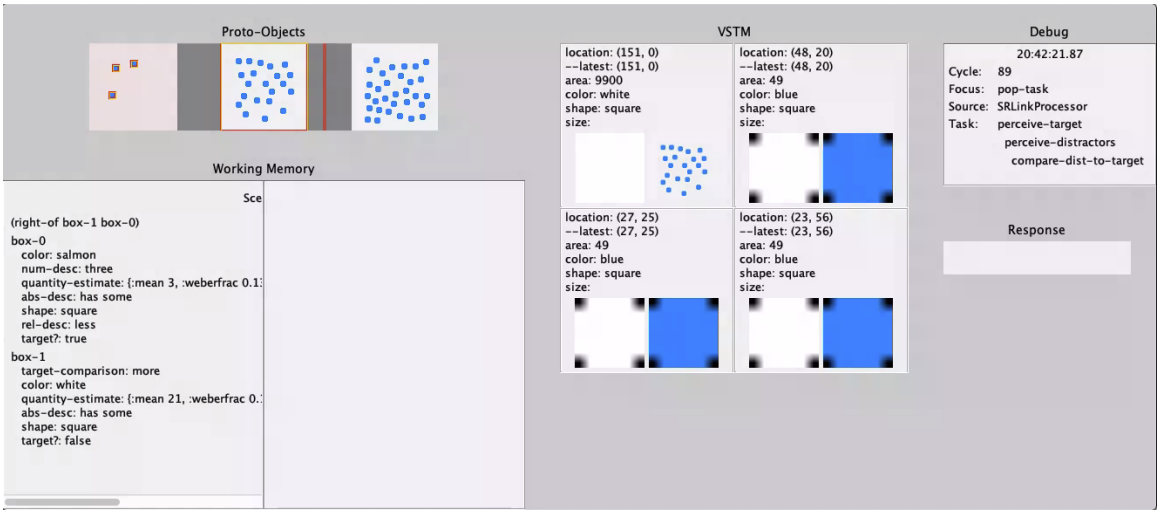

Figure 2: A screenshot of the ARCADIA model of QRE production in mid-run. The vertically oriented red bar in the task image indicates the current scope of spatial attention as it is being swept through the image. The Working Memory display shows the current state of the symbolic scene representation. Note that the last box has not yet been attended and encoded into working memory. Also note that no exact count was obtained for `box-1`, unlike `box-0`, as the quantity estimate indicated that it was too high effort to enumerate exactly.

the number of items n in a group that receives attentional focus, with a standard deviation, $\sigma = w * n$, where w is the Weber fraction associated with the simulated approximate number system. Here $w = 0.13$, based on empirical data (Odic et al., 2013). Exact numbers are represented by count words, and the correct sequence of count words is represented in the system as assumed knowledge.

3.2 Process

Below we give a high-level description of how our model operates. A screenshot of the model in operation can be found in Figure 2. Videos of the model operating over example problems can also be found online.[1]

3.2.1 Incremental Perception

The model performs a left-to-right sweep of spatial attention. For each lightly colored square box it encounters, it encodes into the symbolic scene representation in working memory the existence of a new box of items. Likewise for each of these boxes, spatial attention is then focused on the box and a subtask that encodes information about each collection within the box is initiated. The steps of this subtask are described below.

Step 1: Target/Distractor Classification: The model determines whether or not the current collection is the target collection by considering the background color of the box.

Step 2: Quantity Estimation: The model then attends to the group of items within the box, resulting in an quantity estimate that is encoded and associated with the collection. This quantity estimate is encoded into working memory. Information is then encoded about whether or not items are present or absent (`abs-desc`) from the box.

Step 3: Countability Judgment: Next, a number is randomly sampled from the approximate distribution encoded during the previous quantity estimation step. If this number is lower than a countability threshold value τ_{count}, then the quantity is deemed low effort enough to exactly enumerate and the model proceeds to Step 4. Otherwise, the model determines that exact enumeration will be too high effort, and skips Step 4. Evidence from the literature on numerosity perception provides support for this notion of a countability judgment guiding enumeration strategy (Mandler and Shebo, 1982). Furthermore, we set our countability threshold value $\tau_{count} = 7$ based on experimental results from (Mandler and Shebo, 1982).

Step 4: Exact Enumeration: In the case when a collection is gauged to be countable, an exact

[1] https://osf.io/6rsg7/?view_only=034f98a2449243e28e2a593797039093

count is associated with the box (num-desc). Spatial attention is then swept downwards within the collection. Individual objects are encoded into visual short term memory, and for each new object detected, the count is incremented.

Finally, after all applicable approximate and exact quantity representations are encoded for each box, the quantity information of each distractor is then compared with that of the target to ascertain information about the relative quantity of the target (rel-desc) compared with the distractors. Specifically, exact quantities are first compared if available. In the absence of exact counts, the approximate representations are compared. This is achieved by considering the Gaussian distribution that results from subtracting the Gaussian representing the distractor estimate from the Gaussian representing the target estimate. If the cumulative probability distribution of this new distribution for negative values above a threshold τ_{rel}, this indicates that the target has a smaller quantity. Conversely, if the cumulative probability distribution for positive values is above the threshold, this indicates that the target has a larger quantity. In the scope of this paper, we set $\tau_{rel} = 0.75$. By comparing the target to each distractor in this manner, we determine whether the target is the collection with the most or fewest items (or neither).

3.2.2 Content Selection

Having completed the process of incrementally constructing a symbolic scene representation, the model then begins the REG process. To do this we use a modified version of the the Incremental Algorithm (Dale and Reiter, 1995). The modification is based on the one proposed by Briggs and Harner (2019), wherein attributes with missing values are skipped. To evaluate the model, we used the best performing preference order from (Briggs and Harner, 2019): num-desc > abs-desc > rel-desc.

3.3 Evaluation

As a proof-of-concept evaluation, we ran the model over images corresponding to the twenty unique QRE task problems modeled by Perceptual Cost Pruning by Briggs and Harner (2019). We then compared the resulting output attribute selections to the output reported for the Perceptual Cost Pruning algorithm configured with the same attribute preference order, verifying that it was the same for all problems. The problem descriptions and corresponding model output is found in Table 1.

Problem ID	Target Quantity	Distractor Quantities	Model Output
1	31	{11,11}	{REL}
2	31	{3,3}	{REL}
3	11	{31,31}	{REL}
4	3	{1,1}	{NUM}
5	1	{3,3}	{NUM}
6	3	{31,31}	{NUM}
7	31	{21,11}	{REL}
8	31	{11,3}	{REL}
10	11	{21,31}	{REL}
11	3	{2,1}	{NUM}
13	1	{2,3}	{NUM}
14	3	{21,31}	{NUM}
15	21	{11,0}	{REL,ABS}
16	11	{21,0}	{REL,ABS}
17	3	{1,0}	{NUM}
18	1	{3,0}	{NUM}
19	11	{0,0}	{ABS}
20	12	{0,0}	{ABS}
21	3	{0,0}	{NUM}
22	4	{0,0}	{NUM}

Table 1: QRE problems and resulting model output.

4 Discussion

People can not mentally represent complex visual scenes down to the exact detail. Evidence suggests that people strategically attend to scenes to collect enough detail to complete their task. In other words, people build up sparse scene representations. Successfully modeling human performance on the QRE task requires the ability to model what pieces of information people do or do not encode. Prior work successfully modeled QRE generation through a *destructive* process, in which a complete scene representation is reduced according to models of perceptual cost. Here, we have presented a *constructive* approach to modeling QRE generation by simulating a process of incremental scene perception.

4.1 Approach Advantages

What does the constructive, incremental approach give us over the destructive one? First, it provides a model that begins with raw images, rather than a symbolic knowledge base. Second, we argue that the constructive approach is more extensible when applied to a wider range of possible visual description or REG tasks. The destructive approach requires the development of a separate model of perceptual cost for each attribute that it considers, which for complex domains could necessitate quite

complicated collections of models. In the constructive approach that we present, considerations of perceptual cost and the order in which facts are encoded are built into the perceptual process and how attention is strategically controlled.

One example where a constructive approach involving incremental scene representation would be beneficial is in describing more complex visual scenes with grouped items. In contrast to the stimuli in the present work, which are randomly scattered collections of visual items, common visual scenes often contain items that are organized together in multiple groups. In some cases, visual grouping can reduce the perceptual cost of determining the total quantity of items. Visual grouping allows for people to attend to and enumerate each group separately, establishing the total number through mental arithmetic, instead of slowly counting each item one-by-one (Starkey and McCandliss, 2014; Ciccione and Dehaene, 2020). Recent work indicates that when generating quantified descriptions of scenes with visual groups of the same cardinality, people have a tendency to omit descriptions of total quantity, and instead describe the number of groups and the number of items within each group (Briggs et al., 2020). This finding is consistent with the idea of incremental scene representation and that knowledge about the number of groups and cardinality of each group precedes knowledge of total quantity.

4.2 Future Work

We intend to use the model presented in this paper as a basis to explore how incremental perception can explain individual variation in the forms of QREs observed in (Barr et al., 2013). While the present model waits for the scene representation to be encoded before content generation is started, it could be modified to begin the content planning phase earlier, before the scene is fully processed.

We predict that in certain QRE problems, the order in which distractors are perceived and compared with the target may affect the content of generated expressions. Specifically, consider Problem 16 described in Table 1. The model predicts a QRE containing both a relative and absolute description of quantity. Examples of different expressions that fit this template are: (A) "the one with some, but not the most"; and (B) "the one with fewer, but not the empty one". Expressions similar to forms A and B were found when the human data were reexamined (Briggs and Harner, 2019). We predict that the order in which the two distractors are perceived would determine which expression form is more commonly produced. Form A would correspond to the empty distractor being perceived first, while form B would correspond to the empty distractor being perceived second.

Acknowledgments

This work was supported by an NRL Karles Fellowship awarded to author and AFOSR MIPR grant F4FGA07074G001. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Dale Barr, Kees van Deemter, and Raquel Fernández. 2013. Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 157–161.

Hilary Barth, Nancy Kanwisher, and Elizabeth Spelke. 2003. The construction of large number representations in adults. *Cognition*, 86(3):201–221.

Will Bridewell and Paul F Bello. 2016. A Theory of Attention for Cognitive Systems. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems*, pages 1–16, Evanston, USA.

Gordon Briggs and Hillary Harner. 2019. Generating quantified referring expressions with perceptual cost pruning. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 11–18, Tokyo, Japan.

Gordon Briggs, Hillary Harner, and Sangeet Khemlani. 2020. Visual grouping and pragmatic constraints in the generation of quantified descriptions. In *Proceedings of the 42nd Annual Virtual Meeting of the Cognitive Science Society*, pages 1008–1014.

Lorenzo Ciccione and Stanislas Dehaene. 2020. Grouping mechanisms in numerosity perception.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. 2004. Core systems of number. *Trends in Cognitive Sciences*, 8:307–314.

Rochel Gelman and Charles R Gallistel. 1986. *The child's understanding of number*. Harvard University Press.

Daniel C Hyde and Justin N Wood. 2011. Spatial attention determines the nature of nonverbal number representation. *Journal of Cognitive Neuroscience*, 23:2336–2351.

Andrew Lovett, Will Bridewell, and Paul Bello. 2019. Selection enables enhancement: An integrated model of object tracking. *Journal of Vision*, 19(14):1–31.

George Mandler and Billie J Shebo. 1982. Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, 111:1–22.

Darko Odic, Melissa E Libertus, Lisa Feigenson, and Justin Halberda. 2013. Developmental change in the acuity of approximate number and area representations. *Developmental psychology*, 49(6):1103.

Gillian S Starkey and Bruce D McCandliss. 2014. The emergence of "groupitizing" in children's numerical cognition. *Journal of Experimental Child Psychology*, 126:120–137.

Kees Van Deemter. 2016. *Computational models of referring: a study in cognitive science*. MIT Press.

Alfred L Yarbus. 1967. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer.