ReferItGame: Referring to Objects in Photographs of Natural Scenes

Sahar Kazemzadeh[1*] **Vicente Ordonez**[1*] **Mark Matten**[2] **Tamara L. Berg**[1]
[1]University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2]The Bishop's School, San Diego, CA 92037, USA
vicente@cs.unc.edu,tlberg@cs.unc.edu

Abstract

In this paper we introduce a new game to crowd-source natural language referring expressions. By designing a two player game, we can both collect and verify referring expressions directly within the game. To date, the game has produced a dataset containing 130,525 expressions, referring to 96,654 distinct objects, in 19,894 photographs of natural scenes. This dataset is larger and more varied than previous REG datasets and allows us to study referring expressions in real-world scenes. We provide an in depth analysis of the resulting dataset. Based on our findings, we design a new optimization based model for generating referring expressions and perform experimental evaluations on 3 test sets.

1 Introduction

Much of everyday language and discourse concerns the visual world around us, making understanding the relationship between objects in the physical world and language describing those objects an important challenge problem for AI. From robotics, to image search, to situated language learning, and natural language grounding, there are a number of research areas that would benefit from a better understanding of how people refer to physical entities in the world.

Recent advances in automatic computer vision methods have started to make technologies for recognizing thousands of object categories a near reality (Perronnin et al., 2012; Deng et al., 2012; Deng et al., 2010; Krizhevsky et al., 2012). As a result, there has been a spurt of recent work trying to estimate higher level semantics, including exciting efforts to automatically produce natural language descriptions of images and video (Farhadi et al., 2010; Kulkarni et al., 2011; Yang et al., 2011; Ordonez et al., 2011; Kuznetsova et al., 2012; Feng and Lapata, 2013). Common challenges encountered in these pursuits include the fact that descriptions can be highly task dependent, open-ended, and difficult to evaluate automatically.

Therefore, we look at the related, but more focused problem of referring expression generation (REG). Previous work on REG has made significant progress toward understanding how people generate expressions to refer to objects (a recent survey of techniques is provided in Krahmer and van Deemter (2012)). In this paper, we study the relatively unexplored setting of how people refer to objects in *complex photographs of real-world cluttered scenes*. One initial stumbling block to examining this scenario is lack of existing relevant datasets, as previous collections for studying REG have used relatively focused domains such as graphics generated objects (van Deemter et al., 2006; Viethen and Dale, 2008), crafts (Mitchell et al., 2010), or small everyday (home and office) objects arrayed on a simple background (Mitchell et al., 2013a; FitzGerald et al., 2013).

In this paper, we collect a new large-scale corpus, currently containing 130,525 expressions, referring to 96,654 distinct objects, in 19,894 photographs of real world scenes. Some examples from our dataset are shown in Figure 5. To construct this corpus efficiently, we design a new two player referring expression game (ReferItGame) to crowd-source the data collection. Popularized by efforts like the ESP game (von Ahn and Dabbish, 2004) and Peekaboom (von Ahn et al., 2006b), Human Computation based games can be an effective way to engage users and collect large amounts of data inexpensively. Two player games can also automate verification of human provided annotations.

Our resulting corpus is both more real-world and much bigger than previous datasets, allowing

[*]Indicates equal author contribution.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 787–798,
October 25-29, 2014, Doha, Qatar.

us to examine referring expression generation in a new setting at large scale. To understand and quantify this new dataset, we perform an extensive set of analyses. One significant difference from previous work is that we study how referring expressions vary for different categories. We find that an object's category greatly influences the types of attributes used in their referring expression (e.g. people use color words to describe cars more often than mountains). Additionally, we find that references to an object are sometimes made with respect to other nearby objects, e.g. "the ball to left of the man". Interestingly, the types of reference objects (i.e. "the man") used in referring expressions is also biased toward some categories. Finally, we find that the word used to refer to the object category itself displays consistencies across people. This notion is related to ideas of entry-level categories from Psychology (Rosch, 1978).

Given these findings, we propose an optimization model for generating referring expressions that jointly selects which attributes to include in the expression, and what attribute values to generate. This model incorporates both visual models for selecting attribute-values and object category specific priors. Experimental evaluations indicate that our proposed model produces reasonable results for REG.

In summary, contributions of our paper include:

- A two player online game to collect and verify natural language referring expressions.
- A new large-scale dataset containing natural language expressions referring to objects in photographs of real world scenes.
- Analyses of the collected dataset, including studying category-specific variations in referring expressions.
- An optimization based model to generate referring expressions for objects in real-world scenes with experimental evaluations on three labeled test sets.

The rest of the paper is organized as follows. First we outline related work from the vision and language communities (§2). Then we describe our online game for collecting referring expressions (§3) and provide an analysis of our new ReferItGame Dataset (§4). Finally, we present and evaluate our model for generating referring expressions (§5) and discuss conclusions and future work (§6).

2 Related Work

Referring Expression Generation: There has been a long history of research on understanding how people generate referring expressions, dating back to the 1970s (Winograd, 1972). One common approach is the Incremental Algorithm (Dale and Reiter, 1995; Dale and Reiter, 2000) which uses logical expressions for generation. Much work in REG follows the Gricean maxims (Grice, 1975) which provide principles for how people will behave in conversation.

Recently, there has been progress examining other aspects of the referring expression problem such as understanding what types of attributes are used (Mitchell et al., 2013a), modeling variations between speakers (Viethen and Dale, 2010; Viethen et al., 2013; Van Deemter et al., 2012; Mitchell et al., 2013b), incorporating visual classifiers (Mitchell et al., 2011), producing algorithms to refer to object sets (Ren et al., 2010; FitzGerald et al., 2013), or examining impoverished perception REG (Fang et al., 2013). A good survey of work in this area is provided in Krahmer and van Deemter (2012). We build on past work, extending models to generate attributes jointly in a category specific framework.

Referring Expression Datasets: Some initial datasets in REG used graphics engines to produce images of objects (van Deemter et al., 2006; Viethen and Dale, 2008). Recently more realistic datasets have been introduced, consisting of craft objects like pipecleaners, ribbons, and feathers (Mitchell et al., 2010), or everyday home and office objects such as staplers, combs, or rulers (Mitchell et al., 2013a), arrayed on a simple background. These datasets helped moved referring expression generation research into the domain of real world objects. We seek to further these pursuits by constructing a dataset of natural objects in photographs of the real world.

Image & Video Description Generation: Recent research on automatic image description has followed two main directions. Retrieval based methods (Aker and Gaizauskas, 2010; Farhadi et al., 2010; Ordonez et al., 2011; Feng and Lapata, 2010; Feng and Lapata, 2013) retrieve existing captions or phrases to describe a query image. Bottom up methods (Kulkarni et al., 2011; Yang et al., 2011; Yao et al., 2010) rely on visual classifiers to first recognize image content and then construct captions from scratch, perhaps with some

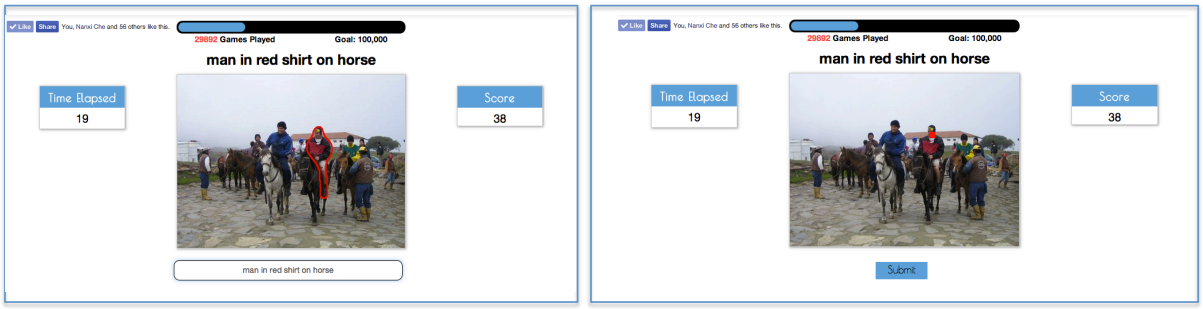


Figure 1: An example game. Player 1 (*left*) sees an image with an object outlined in red (the man) and provides a referring expression for the object ("man in red shirt on horse"). Player 2 (*right*) sees the image and the expression from Player 1 and must localize the correct object by clicking on it (click indicated by the red square). Elapsed time and current scores are also provided.

input from natural language statistics. Very recently, these ideas have been extended to produce descriptions for videos (Guadarrama et al., 2013; Barbu et al., 2012). Like these methods, we generate descriptions for natural scenes, but focus on referring to particular objects rather than providing an overall description of an image or video.

Human Computation Games: Games can be a useful tool for collecting large amounts of labeled data quickly. Human Computation Games were first introduced by Luis von Ahn in the ESP game (von Ahn and Dabbish, 2004) for image labeling, and later extended to segment objects (von Ahn et al., 2006b), collect common-sense knowledge (von Ahn et al., 2006a), or disambiguate words (Seemakurty et al., 2010). Recently, crowd games have also been introduced into the computer vision community for tasks like fine grained category recognition (Deng et al., 2013). These games can be released publicly on the web or used on Mechanical Turk to enhance and encourage turker participation (Deng et al., 2013). Inspired by the success of previous games, we create a game to collect and verify natural language expressions referring to objects in natural scenes.

3 Referring Expression Game (ReferItGame)

In this section we describe our referring expression game (ReferItGame*), a simple two player game where players alternate between generating expressions referring to objects in images of natural scenes, and clicking on the locations of described objects. An example game is shown in Figure 1.

*Available online at http://referitgame.com

3.1 Game Play

Player 1: is shown an image with an object outlined in red and provided with a text box in which to write a referring expression. *Player 2:* is shown the same image and the referring expression written by Player 1 and must click on the location of the described object (note, Player 2 does not see the object segmentation). If Player 2 clicks on the correct object, then both players receive game points and the Player 1 and Player 2 roles swap for the next image. If Player 2 does not click on the correct object then no points are received and the players remain in their current roles.

This provides us with referring expressions for our dataset and verification that the expressions are valid since they led to correct object localizations. Expressions written for games where the object was not correctly localized are kept and released with the dataset for future study, but are not included in our final dataset analyses or statistics. A game timer encourages players to write expressions quickly, resulting in more natural expressions. Also, IP addresses are filtered to prevent people from simultaneously playing both roles.

3.2 Playing Against the Computer

To promote engagement, we implement a single player version of the game. When a player connects, if there is another player online then the two people are paired. If there are currently no other available players, then the person plays a "canned" game against the computer. If at any point another person connects, the canned game ends and the player is paired with the new person.

To implement canned games we seed the game with 5000 pre-recorded referring expression games (5 referring expressions and resulting clicks

for each of 1000 objects) collected using Amazon's Mechanical Turk service. Implementing an automated version of Player 1 is simple; we just show the person one of the pre-collected referring expressions and they click as usual.

Automating the role of Player 2 is a bit more complicated. In this case, we compare the person's written expression against the pre-recorded expressions for the same object. For this comparison we use a parser to lemmatize the words in an expression and then compute cosine similarity between expressions with a bag of words representation. Based on this measure the closest matching expression is determined. If there is no similarity between the newly generated expression and the canned expressions, the expression is deemed incorrect and a random click location (outside of the object) is generated. If there is a successful match with a previously generated expression, then the canned click from the most similar pre-recorded game is used. More complex similarities could be used, but since we require real-time performance in our game setting we use this simple implementation which works well for our expressions.

4 ReferItGame Dataset

In this section we describe the ReferItGame dataset[†], including images and labels, processing the dataset, and analysis of the collection.

4.1 Images and Labels

We build our dataset of referring expressions on top of the ImageCLEF IAPR image retrieval dataset (Grubinger et al., 2006). This dataset is a collection of 20,000 images available free of charge without copyright restrictions, depicting a variety of aspects of everyday life, from sports, to animals, to cities, and landscapes. Crucial for our purposes, the SAIAPR TC-12 expansion (Escalante et al., 2010) includes segmentations of each image into regions indicating the locations of constituent objects. 238 different object categories are labeled, including animals, people, buildings, objects, and background elements like grass or sky. This provides us with information regarding object category, object location, and object size, as well as the location and categories of other objects present in the same image.

[†] Available at http://tamaraberg.com/referitgame

4.2 Collecting the Dataset

From the ImageCLEF dataset, we created a total of over 100k distinct games (one per object labeled in the dataset). For the games we imposed an ordering to allow for collecting the most interesting expressions first. Initially we prioritized games for objects in images with multiple objects of the same category. Once these games were completed, we prioritized ordering based on object category to include a comprehensive range of objects. Finally, after successfully collecting referring expressions from the prioritized games, we posted games for the remaining objects. In order to evaluate consistency of expression generation across people, we also include a probability of repeating previously played games during collection.

To date, we have collected 130,525 successfully completed games. This includes 10,431 canned games (a person playing against the computer, not including the initial seed set) and 120,094 real games (two people playing). 96,654 distinct objects from 19,984 photographs are represented in the dataset. This covers almost all of the objects present in the IAPR corpus. The remaining objects from the collection were either too small or too ambiguous to result in successful games.

For data collection, we posted the game online for anyone on the web to play and encouraged participation through social media and the survey section of reddit. In this manner we collected over 4 thousand referring expressions over a period of 3 weeks. To speed up data collection, we also posted the game on Mechanical Turk. Turkers were paid upon completion of 10 correct games (games where Player 2 clicks on the correct object of interest). Turkers were pre-screened to have approval ratings above 80% and to be located in the US for language consistency.

4.3 Processing the Dataset

Because of the size of the dataset, hand annotation of all referring expressions is prohibitive. Therefore, similar to past work (FitzGerald et al., 2013), we design an automatic method to pre-process the expressions and extract object and attribute mentions. These automatically processed expressions are used only for analysis and model training. We also fully hand label portions of the dataset for evaluation (§5.2).

By examining the expressions in the collected dataset, we define a set of attributes with broad

$$
\begin{aligned}
S &::= subject\_word \\
color\_word' &::= rel(S, color\_word)_{color\_word'=color\_word} \mid \\
&\quad prep\_in(S, color\_word)_{color\_word'=color\_word} \\
size\_word' &::= rel(S, size\_word)_{size\_word'=size\_word} \\
abs\_loc\_word' &::= rel(S, abs\_loc\_word)_{\ abs\_loc\_word'=abs\_loc\_word} \mid \\
&\quad prep\_on(S, orientation\_word) \wedge \neg prep\_of(S, \_)_{abs\_loc\_word'=on+orientation\_word} \\
rel\_loc\_word' &::= RL \\
RL &::= prep\_rel\_loc\_word(S, object\_word)_{RL=rel\_loc\_word} \mid \\
&\quad prep\_on(S, orientation\_word) \wedge prep\_of(S, object\_word)_{\ RL=on\_orientation\_word} \mid \\
&\quad prep\_to(S, orientation\_word) \wedge prep\_of(S, object\_word)_{\ RL=to\_orientation\_word} \mid \\
&\quad prep\_at(S, orientation\_word) \wedge prep\_of(S, object\_word)_{\ RL=at\_orientation\_word} \\
generic\_word' &::= amod(S, generic\_word)
\end{aligned}
$$

Figure 2: Templates for parsing attributes from referring expressions (§4.3).

coverage of the attribute types used in the referring expressions. We define the set of attributes for a referring expression as a 7-tuple $R = \{r_1, r_2, r_3, r_4, r_5, r_6, r_7\}$:

- r_1 is an entry-level category attribute,
- r_2 is a color attribute,
- r_3 is a size attribute,
- r_4 is an absolute location attribute,
- r_5 is a relative location relation attribute,
- r_6 is a relative location object attribute,
- r_7 is a generic attribute,

Color and *size* attributes refer to the object color (e.g. "blue") and object size (e.g. "tiny") respectively. *Absolute location* refers to the location of the object in the image (e.g. "top of the image"). *Relative location relation* and *relative location object* attributes allow for referring expressions that localize the object with respect to another object in the picture (e.g. "the car to the left of the tree"). *Generic attributes* cover all less frequently observed attribute types (e.g. "wooden" or "round").

The *entry-level category attribute* is related to the concept of entry-level categories first proposed by Psychologists in the 1970s (Rosch, 1978) and recently explored in visual recognition (Ordonez et al., 2013). The idea of entry-level categories is that an object can belong to many different categories; an indigo bunting is an oscine, a bird, a vertebrate, a chordate, and so on. But, a person looking at a picture of one would probably call it a bird (unless they are very familiar with ornithology). Therefore, we include this attribute to capture how people name object categories in referring expressions.

Parsing the referring expressions: We parse the expressions using the most recent version of the StanfordCoreNLP parser (Socher et al., 2013). We begin by traversing the parse tree in a breadth-first manner and selecting the head noun of the sentence to determine the object of the referring expression, denoted as $subject\_word$. We pre-define a dictionary of attribute-values ($color\_word$, $size\_word$, $abs\_location\_word$, $rel\_location\_word$) for each of the attributes based on the observed data using a combination of POS-tagging and manual labeling.

We then apply a template-based approach on the collapsed dependency relations to recover the set of attributes (the main template rules are shown in Figure 2). The relationship rel indicates any linguistic binary relationship between the subject word S and another word, including the $amod$ relationship. $Orientation\_word$ captures the words like left, right, top and bottom. For $generic\_word$ we consider any modifier words other than those captured by our other attributes (color, size, location).

Using this template-based parser we can for instance parse the following expression: "Red flower on top of pedestal". The first rule would match the $prep(S, color\_word)$ relation, effectively recovering the attribute $color\_word'$ as "red". The second rule would match the $prep\_on(S, orientation\_word) \wedge prep\_of(S, object\_word)$ relations, recovering rel_loc_word' as "on top of " and $object\_word$ as "pedestal".

The accuracy of our parser based processing is 91%. This was evaluated on 4,500 expressions

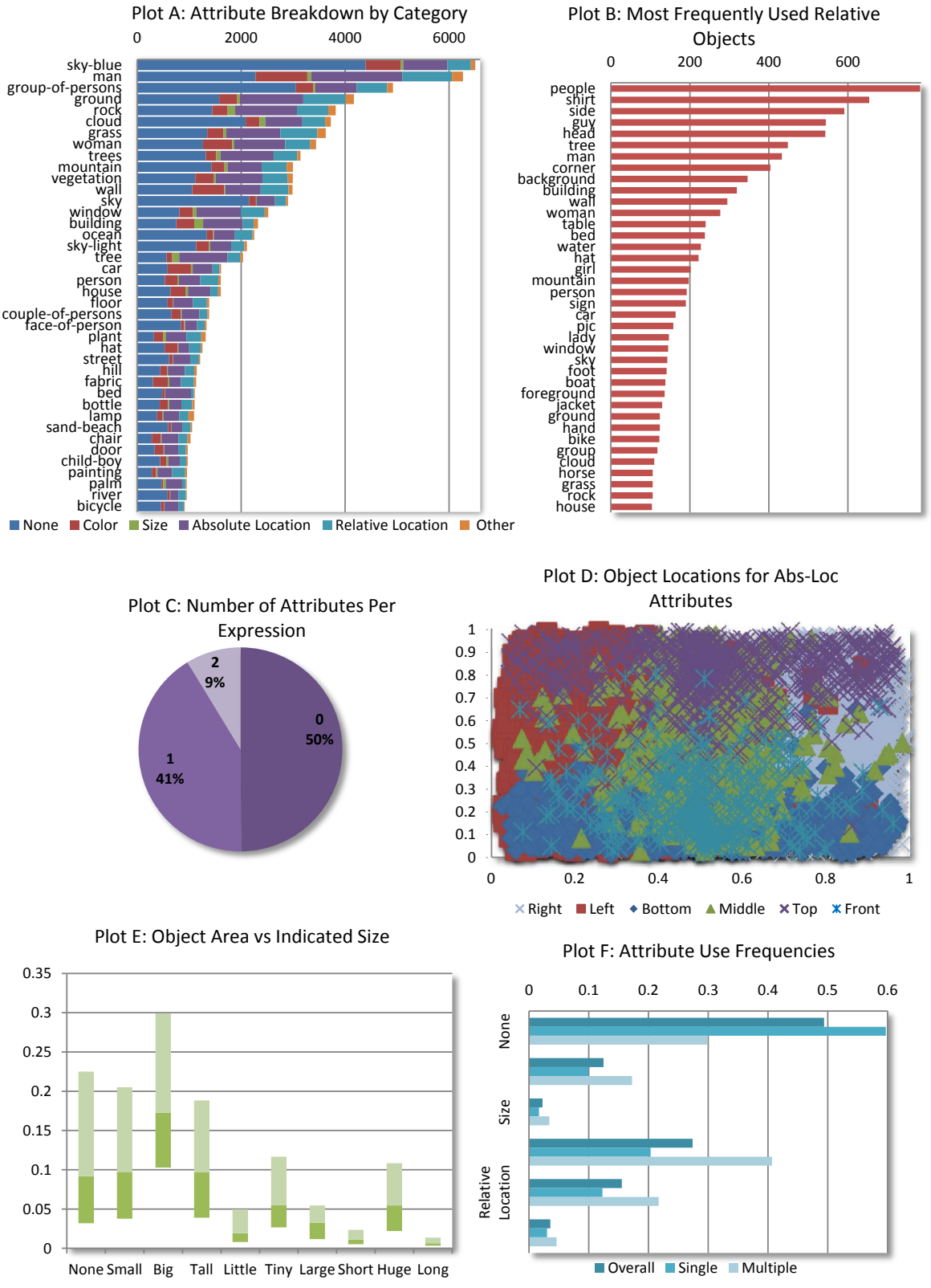


Figure 3: Analyses of the ReferItGame Dataset. **Plot A** shows frequency and attribute occurrence for common object categories. **Plot B** shows objects frequently used as reference points, ie "to the left of the man". **Plot C** shows frequencies of using 0, 1 or 2 attributes within the same expression. **Plot D** shows object locations vs location words used. **Plot E** shows normalized object size vs size words used (bars show 1^{st} through 3^{rd} quartiles). **Plot F** shows the frequency of usage of each attribute type for images containing either a *single* instance of the object category or *multiple* instances of the category.

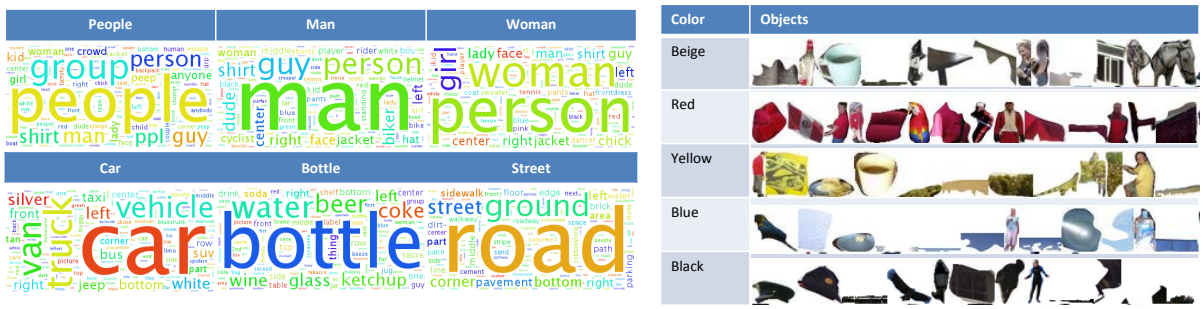


Figure 4: **Left:** Tag clouds showing entry-Level category words used in referring expressions to name various object categories, with word size indicating frequency. For example, this indicates that "streets" are often called "road", sometimes "ground", sometimes "roadway", etc. **Right:** example objects predicted to portray some of our color attribute values. Note sometimes our color predictor is quite accurate, and sometimes it makes mistakes (see the man in a red shirt predicted as "yellow").

that were manually parsed by a human annotator.

4.4 Dataset Analysis

In the resulting dataset, we have a range of coverage over objects. For 10,304 of the objects we have 2 or more referring expressions while for the rest of the objects we have collected only one expression. This creates a dataset that emphasizes breadth while also containing enough data to study speaker variation.

Multiple attribute analyses are provided in Figure 3. We find that most expressions use 0, 1, or 2 attributes (in addition to the entry-level attribute object word), with very few expressions containing more than 2 attributes (frequencies are shown in Fig 3c). We also examine what types of attributes are used most frequently, according to object category in Fig 3a, and when associated with single or multiple occurrences of the same object category in an image in Fig 3f. The frequency of attribute usage in images containing multiple objects of the same type increases for all types, compared to single object occurrences. Perhaps more interestingly, the use of different attributes is highly category dependent. People use more attribute words overall to describe some categories, like "man", "woman", or "plant", and the distribution of attribute types also varies by category. For example, color attributes are used more frequently for categories like "car" or "woman" than for categories like "sky" or "rock".

We also examine which objects are most frequently used as points of reference, e.g.,"the chair next to the *man*" in Fig 3b. We observe that people and some background categories like "tree" or "wall" are often used to help localize objects in referring expressions. Additionally, we provide plots showing the relationship between object location in the image and use of absolute location words, Fig 3d, as well as size words vs object area, Fig 3e.

Finally, we study entry-level category attribute-values to understand how people name objects in referring expressions. Tag clouds indicating the frequencies of words used to name various object categories are provided in Fig 4 (left). Objects like "street" are usually referred to as "road", but sometimes they are called "ground", "roadway", etc. "Bottles" are usually called "bottle", but sometimes referred to as "coke" or "beer". Interestingly, "man" is usually called "man" while "woman" is most often called "person" in the referring expressions.

5 Generating Referring Expressions

In this section we describe our proposed generation model and provide experimental evaluations on three test sets.

5.1 Generation Model

Given an input tuple $I = \{P, S\}$, where P is a target object and S is a scene (image containing multiple objects), our goal is to generate an output referring expression, R. For instance, the representation R for the referring expression: *The big old white cabin beside the tree* would be $R = \{cabin, white, big, \varnothing, beside, tree, old\}$.

To generate referring expressions we construct vocabularies V_{r_i} with candidate values for each attribute $r_i \in R$, where attribute vocabulary V_{r_i} contains the set of words observed in our parsed referring expressions for attribute r_i plus an additional

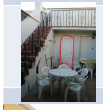
Image	Human Expressions	Generated Expressions
	picture on the wall picture picture	**Baseline:**[picture, white, , right, , ,] **Full:** [picture, , , , prep_on, wall,]
	Door white door middle white door	**Baseline:**[door, white, , right, , ,] **Full:**[door, white, , right, , ,]
	big gated window on right of white section black big window right brown railings on right	**Baseline:**[window, white, , right, , ,] **Full:**[window, brown, , right, , ,]
	white shirt man white shirt on right man on right	**Baseline:**[man, white, , right, , ,] **Full:**[man, white, , right, , ,]
	building on right behind guys blue right building building on right	**Baseline:**[building, white, , right, , ,] **Full:**[building, white, , right, , ,]

Image	Human Expressions	Generated Expressions
	picture santa the santa picture	**Baseline:**[picture, white, , right, , ,] **Full:**[picture, , , , prep_on, plant,]
	right doorway right brown door right door	**Baseline:**[door, , , right, prep_on, person,] **Full:**[door, , , right, prep_above, person,]
	with flag window top 2nd left 2nd window top left	**Baseline:**[window, , , right, prep_on, person,] **Full:**[window, , , left, prep_above, door,]
	red guy left sitting left bottom guy red shirt lef	**Baseline:**[man, , , right, prep_on, wall,] **Full:**[man, , , left, prep_in, woman,]
	buildings buildings buildings	**Baseline:**[building, white, , right, , ,] **Full:**[building, brown, , middle, , ,

Figure 5: Example results, including human generated expressions, baseline and full model generated expressions. For some images the model does well at mimicking human expressions (left). For others it does not generate the correct attributes (right).

ε value indicating that the attribute should be omitted from the referring expression entirely.

In this way, our framework can jointly determine which attributes to include in the expression (e.g.,"size" and "color") and what attribute values to generate (e.g.,"small" and "blue") from the list of all possible values. We enforce a constraint to always include an "entry-level category" attribute (e.g. "boy") so that we always generate a word referring to the object.

We pose our problem as an optimization where we map a tuple $\{P, S\}$ to a referring expression R^* as:

$$R^* = \underset{R}{\text{argmax}} E(R, P, S) \quad \text{s. t. } f_i(R) \leq b_i \tag{1}$$

Where the objective function E is decomposed as:

$$E(R,P,S) = \alpha \sum_{i=2}^{6} \phi_i(r_i, P, S) + \beta \sum_{i=1}^{7} \psi_i(r_i, type(P)) + \sum_{i>j} \psi_{i,j}(r_i, r_j) \tag{2}$$

Where ϕ_i is the compatibility function between an attribute-value for r_i and the properties of the observed scene S and object P (described in §5.1.1). The terms ψ_i and $\psi_{i,j}$ are unary and pairwise priors computed based on observed co-occurrence statistics of attribute-values for r_i with categories (where $type(P)$ denotes the type or category of an object) and between pairs of attribute-values (described in §5.1.2). Attributes r_1 and r_7 are modeled only in the priors since we do not have visual models for these attributes.

The constraints $f_i(R) \leq b_i$ are restricted to be linear constraints and are used to impose hard constraints on the solution. The first such constraint is used to control the verbosity (length) of the generated referring expression using a constraint function that imposes a minimum attribute length requirement by restricting the number of entries r_i that can take value ε in the solution.

$$\sum_i \mathbb{1}[r_i = \varepsilon] \leq 7 - \gamma(P, S) \tag{3}$$

Where $\mathbb{1}[.]$ is the indicator function and $\gamma(P, S)$ is a term that allows us to change the length requirement based on the object and scene (so that images with a larger number of objects of the same type have a larger length requirement).

Finally we add hard constraints such that $r_5 = \varepsilon \iff r_6 = \varepsilon$, so that relative location and relative object attributes are produced together.

5.1.1 Content-based potentials

Potentials ϕ_i are defined for attributes r_2 to r_6. Attribute r_7 represents a variety of different attributes, e.g. material or shape attributes, but we lack sufficient data to train visual models for these infrequent attribute terms. Therefore, we model these attributes using only prior statistics-based potentials (§5.1.2). Visual recognition models for recognizing entry-level object categories

could also be incorporated for modeling r_1, but we leave this as future work.

Color attribute:

$$\phi_2(r_2 = c_k, P, S) = sim(hist_{c_k}, hist(P))$$

Where $hist(P)$ is the HSV color histogram of the object P. We compute similarity sim using cosine similarity, and $hist_{c_k}$ is the mean histogram of all objects in our training data that were referred to with color attribute-value $c_k \in V_{r_2}$.

Size attribute:

$$\phi_3(r_3 = s_k, P, S) = \frac{1}{\sigma_{s_k}\sqrt{2\pi}} e^{-\left(size(P)-\mu_{s_k}\right)^2 / 2\sigma_{s_k}^2}$$

Where $size(P)$ is the size of object P normalized by image size. We model the probabilities of each size word $s_k \in V_{r_3}$ as a Gaussian learned on our training set.

Absolute-location attribute:

$$\phi_4(r_4 = a_k, P, S) = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{a_k}|}} e^{-\frac{1}{2}(loc(P)-\mu_{a_k})^T \Sigma_{a_k}{}^{-1} (loc(P)-\mu_{a_k})}$$

Where $loc(P)$ are the 2-dimensional coordinates of the object P normalized to be $\in [0-1]$. Parameters μ_{a_k} and Σ_{a_k} are estimated from training data for each absolute location word $a_k \in V_{r_4}$.

Relative-location and Relative object:

$$\phi_5(r_5 = l_k, P, S) = \mathbb{1}[l_k = \varepsilon] \cdot g(count(type(P), S))$$

If there are a larger number of objects of the same type in the image we find that the probability of using a relative-location-object increases (e.g., "the car to the right of the man"). For images where P was the only object of that category type, the probability of using a relative-location-object is 0.12. This increases to 0.22 when there were two objects of the same type and further increases to 0.26 for additional objects of the same type. Therefore, we model the probability of selecting relative location value $l_k \in V_{r_5}$ as a function g, where $count(type(P), S)$ counts the number of objects in the scene S of the same category type as the object P.

$$\phi_6(r_6 = o_k, P, S) = \mathbb{1}[o_k \in objectsnear(location(P), S)]$$

The above expression filters out potential relative objects $o_k \in V_{r_6}$ that are not located in sufficient proximity to object P or are not present in the image at all.

5.1.2 Prior statistics-based potentials

Prior statistics-based potentials are modeled for all of the attributes r_1 - r_7. Note that these potentials do not depend on specific attribute-values but only on the given object category $type(P)$.

Unary prior potentials ψ_i are defined as:

$$\psi_i(r_i, type(P)) = \frac{\sum_{j=1}^{|D|} \mathbb{1}[(r_i^{(j)} \neq \epsilon) \wedge (type(P^{(j)}) = type(P))]}{\sum_{j=1}^{|D|} \mathbb{1}[type(P^{(j)}) = type(P)]} + \frac{\sum_{j=1}^{|D|} \mathbb{1}[r_i^{(j)} \neq \epsilon]}{|D|} + \lambda$$

Where $D = \{P^{(j)}, S^{(j)}, R^{(j)}\}$ is our training dataset and λ is a small additive smoothing term. The two terms in the above expression represent *category-specific* counts and *global* counts of the number of times a given attribute r_i was output in a referring expression in training data. Pairwise prior potentials $\psi_{i,j}$ are defined as:

$$\sum_{i<j} \psi_{i,j}(r_i, r_j) = \sum_{i<j} \psi_{i,j}^{(1)}(r_i, r_j) + \psi_{5,6}^{(2)}(r_5, r_6)$$

$$\psi_{i,j}^{(1)}(r_i, r_j) = \begin{cases} 1 & \text{if } r_i = r_j = \varepsilon \\ C + \lambda & o.w. \end{cases}$$

$$\psi_{5,6}^{(2)}(r_5 = a, r_6 = b) = \frac{\sum_{t=1}^{|D|} \mathbb{1}[(r_5^{(t)} = a) \wedge (r_6^{(t)} = b)]}{|D|}$$

where $C = \frac{\sum_{t=1}^{|D|} \mathbb{1}[(r_i^{(t)} \neq \epsilon) \wedge (r_j^{(t)} \neq \epsilon)]}{|D|}$. The pairwise potential $\psi_{i,j}^{(1)}$ captures the pairwise statistics of how frequently people use pairs of attribute types.

SOURCE	PREC(%)	RECALL(%)
Baseline - A	27.92	43.27
Full Model - A	**36.28**	**53.44**
Baseline - B	29.87	50.57
Full Model - B	**36.68**	**59.80**
Baseline - C	28.85	37.41
Full Model - C	**37.73**	**48.54**

Table 1: Baseline Model & Full Model performance on the three test sets (A,B,C).

For instance how frequently people use both color and size attributes to refer to an object. The pairwise potential $\psi_{i,j}^{(2)}$ produces a cohesion score between relative-location words and relative-object words based on global dataset statistics.

5.2 Experiments

We implement the proposed model using commercial binary integer linear programming software (IBM ILOG CPLEX). This requires introducing a set of indicator variables for each of our multi-valued attributes and another set of indicator variables to model pairwise interactions between our variables, as well as incorporating additional consistency constraints between variables. Model parameters (α and β) are tuned on data randomly sampled from the training set.

Test Sets: We evaluate our model on three test sets, each containing 500 objects. For each object in the test sets we collect 3 referring expressions using the ReferItGame and manually label the attributes mentioned in each expression. We find human agreement to be 72.31% on our dataset (where we measure agreement as mean matching accuracy of attribute values for pairs of users across images in our test sets). The three test sets are created to evaluate different aspects of our data.

Test Set A contains objects sampled randomly from the entire dataset. This test set is meant to closely resemble the full dataset distribution. The goal of the other two test sets is to sample expressions for "interesting" objects. We first identify categories that are mainly related to background content elements, e.g. "sky, ground, floor, sand, sidewalk, etc". We consider these categories to be potentially less interesting for study than categories like people, animals, cars, etc. *Test Set B* contains objects sampled from the most frequently occurring object categories in the dataset, selected to contain a balanced number of objects from each category, excluding the less interesting categories. *Test Set C* contains objects sampled from images that contain at least 2 objects of the same category, excluding the less interesting categories.

Results: *Qualitative examples* are shown in Fig 5 comparing our results to the human produced expressions. For some images (left) we do quite well at predicting the correct attributes and values. For others we do less well (right). We also show example objects predicted for some color words in Fig 4 (right). We see that our model can fail in several ways, such as generating the wrong attribute-value due to inaccurate predictions by visual models or selecting incorrect attributes to include in the generated expression.

Quantitative results: precision and recall measures for the 3 test sets are reported in Table 1, including evaluation of a baseline version of our model which incorporates only the prior potentials (§5.1.2) without any content based estimates. We see that our model performs reasonably on both measures, and outperforms the baseline by a large margin on all test sets, with highest performance on the broadly sampled interesting category test set. Note that our problem is somewhat different than traditional REG where the input is often attribute-value pairs and the task is to select which pairs to include in the expression. Our goal is to jointly select which attributes to include and what values to predict from a list of all possible values for the attribute.

6 Conclusions & Future Work

In this paper we have introduced a new game to crowd-source referring expressions for objects in natural scenes. We have used this game to produce a new large-scale dataset with analysis. We have also proposed an optimization based model for REG and performed experimental evaluations. Future work includes developing fully automatic visual recognition methods for REG in real world scenes, and incorporating linguistically inspired models for entry-level category prediction.

Acknowledgments

This work was funded by NSF Awards #1417991 and #1444234. M.M. was supported by the Stony Brook Simons Summer Research Program for High School students. We also thank Alex Berg for many helpful discussions.

References

Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Association for Computational Linguistics (ACL)*.

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven J. Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, Lara Schmidt, Jiangnan Shangguan, Jeffrey Mark Siskind, Jarrell W. Waggoner, Song Wang, Jinlian Wei, Yifan Yin, and Zhiqi Zhang. 2012. Video in sentences out. In *Uncertainty in Artificial Intelligence (UAI)*.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science (CogSci)*, 19:233264.

Robert Dale and Ehud Reiter. 2000. Building natural language generation systems. In *Cambridge University Press*.

Jia Deng, Alexander C. Berg, Kai Li, and Fei-Fei Li. 2010. What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision (ECCV)*.

Jia Deng, Alex Berg, Sanjeev Satheesh, Hao Su, Aditya Khosla, and Fei-Fei Li. 2012. Large scale visual recognition challenge. In *http://www.image-net.org/challenges/LSVRC/2012/index*.

Jia Deng, Jonathan Krause, and Li Fei-Fei. 2013. Fine-grained crowdsourcing for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hugo Jair Escalante, Carlos A. Hernandez, Jesus A. Gonzalez, A. Lopez-Lopez, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villasenor, and Michael Grubinger. 2010. The segmented and annotated iapr tc-12 benchmark. *Computer Vision and Image Understanding (CVIU)*.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *Empirical Methods on Natural Language Processing (EMNLP)*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: generating sentences for images. In *European Conference on Computer Vision (ECCV)*.

Yansong Feng and Mirella Lapata. 2010. How many words is a picture worth? automatic caption generation for news images. In *Association for Computational Linguistics (ACL)*.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In *Empirical Methods on Natural Language Processing (EMNLP)*.

H. Paul Grice. 1975. Logic and conversation. page 4158.

Michael Grubinger, Paul D. Clough, Henning Muller, and Thomas Deselaers. 2006. The iapr benchmark: A new evaluation resource for visual information systems. In *Proceedings of the International Workshop OntoImage (LREC)*.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *International Conference on Computer Vision (ICCV)*.

Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. In *Computational Linguistics*, volume 38, page 173218.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Babytalk: Understanding and generating simple image descriptions. In *IEEE Computer Vision and Pattern Recognition (CVPR)*.

Polina Kuznetsova, Vicente Ordonez, Alex Berg, Tamara L Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Association for Computational Linguistics (ACL)*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *International Natural Language Generation Conference (INLG)*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2011. Two approaches for generating size modifiers. In *European Workshop on Natural Language Generation*.

Margaret Mitchell, Ehud Reiter, and Kees van Deemter. 2013a. Typicality and object reference. In *Cognitive Science (CogSci)*.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013b. Generating expressions that refer to visible objects. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. From large scale image categorization to entry-level categories. In *International Conference on Computer Vision (ICCV)*.

Florent Perronnin, Zeynep Akata, Zaid Harchaoui, and Cordelia Schmid. 2012. Towards good practice in large-scale learning for image classification. In *Computer Vision and Pattern Recognition (CVPR)*.

Yuan Ren, Kees Van Deemter, and Jeff Z Pan. 2010. Charting the potential of description logic for the generation of referring expressions. In *International Natural Language Generation Conference (INLG)*.

Eleanor Rosch. 1978. Principles of categorization. *Cognition and Categorization*, page 2748.

Nitin Seemakurty, Jonathan Chu, Luis von Ahn, and Anthony Tomasic. 2010. Word sense disambiguation via human computation. In *Human Computation Workshop*.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing With Compositional Vector Grammars. In *Association for Computational Linguistics (ACL)*.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *International Conference on Natural Language Generation (INLG)*.

Kees Van Deemter, Albert Gatt, Roger PG van Gompel, and Emiel Krahmer. 2012. Toward a computational psycholinguistics of reference production. In *Topics in Cognitive Science*, volume 4(2), page 166183.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *International Natural Language Generation Conference (INLG)*.

Jette Viethen and Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Australasian Language Technology Workshop*.

Jette Viethen, Margaret Mitchell, and Emiel Krahmer. 2013. Graphs and spatial relations in the generation of referring expressions. In *European Workshop on Natural Language Generation*.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *ACM Conf. on Human Factors in Computing Systems (CHI)*.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006a. Verbosity: A game for collecting common-sense knowledge. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006b. Peekaboom: A game for locating objects in images. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1191.

Yezhou Yang, Ching Lik Teo, Hal Daume III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Empirical Methods on Natural Language Processing (EMNLP)*.

Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2t: Image parsing to text description. *Proc. IEEE*, 98(8).