Do Objects in Real-World Images Have a Canonical Name?

Anonymous EMNLP-IJCNLP submission

Abstract

Research in language & vision commonly acknowledges the fact that speakers can interpret and describe visual scenes in many different ways. For instance, frameworks in image captioning typically elicit (during collection) or train and test on (during modeling) multiple valid alternatives. Individual visual objects and their names, however, are often collected and modeled in simple annotation set-ups where annotators or systems identify objects via bounding boxes and label them with a single "correct" name. A prime example here is VisualGenome that features complex and varied annotations on the level of scenes and regions, but provides canonicalized objects linked to a single name. In this paper, we put the assumption that objects in complex visual scenes bear a canonical name to an empirical test. Based on pre-annotated object bounding boxes in VisualGenome, we elicit multiple names from 36 subjects per object and investigate the extent to which there is variation in the names chosen by different people for the same object. Our analysis reveals that non-canonicalized object naming data shows a lot of interesting linguistic variation as speakers name objects on different levels of specificity (cow-animal) or verbalize different aspect of the same object (bowl-salad). At the same time, we find that object naming prompted via bounding boxes is subject to a certain amount of noise as speakers have problems re-identifying the object that was annotated by the original bounding box. We investigate whether a state-of-the-art model of object labeling implicitly encodes similar variation in object naming and discuss implications for research in language & vision.

1 Introduction

Expressions describing or referring to objects in visual scenes typically include object names: e.g., *cheesecake* or *dessert* in Figure **??**. Determining these object names is a core aspect of virtually every language & vision task, ranging from e.g. referring expression generation to visual dialogue (**?**). We investigate the extent to which there is variation in the names chosen by different people for the same object, and its implications for research in language & vision.

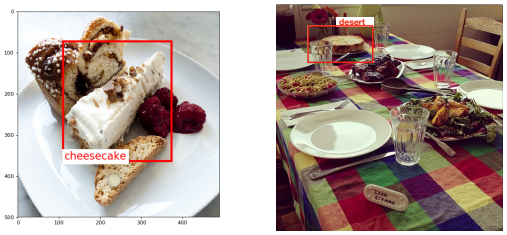

Figure 1: Two objects of the same type of cake, with different names in VisualGenome

Our paper puts together two strands of research that have mostly been pursued independently to date. On the one hand, state-of-the-art computer vision systems are able to accurately classify images into thousands of different categories (e.g. **?**), where the task is often to predict the name for a given object. *//g: Is this true? Imagenet task asks for synsets, which can be taken to be categories... To refine//* However, they mostly adopt very simple assumptions with respect to the underlying lexicon, which is implemented as a simple, flat labeling scheme: A standard object recognition system would be trained to classify the left object in Figure 1 as *cheesecake*, the right one as *dessert*, and using *dessert* for the left picture would be considered incorrect. On the other hand, research on object naming in Cognitive Science has shown that people choose different names depending on the circumstances, with factors such as context or the prototypicality of the object with respect to the category playing a role (**?**). *//g: This research also argues that there is high agreement*

in how people name objects; to do: make coherent.// However, this research typically uses stylized drawings are used, and is focused on taxonomic relations (*sparrow-bird*). *//sz: It is thus unclear how findings from these stylized settings generalize to tasks in language & vision like referring expression generation, where naming is a core aspect. Therefore, in contrast to traditional naming norm studies in Cognitive Science we study object naming in realistic scenes where objects are situated in a natural context! (This comes with additional challenges, like potential object occlusion, background/foreground confusion etc.)//*

In our study, we collect large-scale object naming data via crowdsourcing. Like object naming studies in Cognitive Science, we collect multiple names per object (concretely, 36); like most work on language & vision, we use natural images *//sz: (showing objects in complex visual contexts, surrounded by other objects, not ImageNet-like images)//* on a large scale, annotating objects in 25K images from the Visual Genome dataset. We analyze the agreement in object naming across subjects, and the sources of variation. We find that: *//g: To be put in paragraph form//*

- there is quite a high level of agreement in the task, with the relative frequency of the most common name being 70% on average. This is in accordance with previous results in Cognitive Science (**?**);
- the level of agreement in object naming is much higher in certain domains than in others; as it happens, the domains that have been traditionally used in object naming research (e.g. animals) seem to display the highest amount of agreement in our data set;
- most of the variation in our dataset comes from alternative names that do not stand in a taxonomic relation, suggesting that the previous work in Cognitive Science is missing much of the empirical ground.

  our datasets contains a lot of variability for names coming from different parts of the taxonomy (*dessert* vs. *cake*, *bottle* vs. *wine*)

Moreover, we analyze whether current models implicitly encode the variation in naming, by doing XXX. We find YYY.

2 Data collection

3 Analysis: Agreement

In this Section, we investigate to what extent names annotated in VisualGenome and elicited in ManyNames can be considered canonical, i.e. to what extent speakers agree in their naming choices. Whereas traditional picture naming studies typically use a prototypical image per category and, hence, are mostly interested in the agreement on concept or category-level, we carry out an anaylsis on two different levels: First, we will look at instances and see to what extent names overlap for the same object. Second, we will uses the existing annotation of names in VG to analyze agreement on the level of categories.

3.1 Measures

We compute the following agreement measures:

- **% top**: the average relative frequency of the most frequent response (shown in percent)
- H: the H agreement measure used previously in the psycholinguistic literature
- **N**: the average number of types in the response set of MN
- $\mathbf{N}_{>1}$: the average number of types, excluding types that have been annotated only once
- **top=VG**: the proportion of items where the top response in MN corresponds to the VG name
- **% VG**: the average relative frequency of the VG name in the response set

For measuring **instance-level agreement**, we consider all names annotated for an object as a response set and then average over these response sets. Furthermore, we compute **category-level agreement** by merging the response sets for all objects that have the same VG name and compute the measures over these aggregated response sets.

3.2 Results

Table 1 shows the analysis of the instance-level and category-level agreement. On the instance-level, our annotators achieve a fair amount of overlap in their object naming choices. Thus, for roughly 70% of our objects, the most frequent response in MN corresponds to the original VG

name and, similarly, the average frequency of the top response is also 70%. Generally, this seems to suggest that indeed many objects in our data set have a canonical name. At the same time, the average number of name types per object (5.7 and 2.9, when excluding low-frequency types in each response set) suggests that there is a stable amount of naming variants that can be elicited for instances. Furthermore, the agreement varies quite considerably among domains: in the animal domain, which is often discussed in the object naming literature, annotators achieve a very stable and robust agreement over 90% and an H agreement which comes close to 0 (where 0 is perfect agreement). The people domain, on the other hand, is subject to much more variation and agreement is dramatically lower here, and comes close to 50% for % top.

Finally, the category-level agreement figures tell yet another story: when aggregating the responses for all objects with the same VG name, we obtain on average 28 types (with $n > 1$), i.e. 27 variants of the original VG name. Surprisingly, here, only 29.4% of the aggregated response sets still have the VG name as the most frequent response, which means that for 70% of the VG names, annotators in MN, on average, prefer a different name. Likewise, the relative frequency of the top response drops considerably and H increases from 1.3 for instance-level agreement to 2.4 on object-level agreement. What does this discrepancy between the instance-level and category-level agreement in VG and MN naming choices mean? First of all, it suggests that the same original VG name can trigger very different variants depending on the visual instance, leading to a drastic increase of variants elicited for categories as compared to instances. Second, this clearly shows that annotators in VG do not generally annotate the most canonical name and that many names annotated for objects in VG do not correspond to the overall most preferred variant. *//sz: think more .../*

3.3 Qualitative Analysis

4 Conclusions

	Instance-level agreement							Category-level agreement						
domain	% top	H	N	$N_{>1}$	top=VG	% VG	# Obj	% top	H	N	$N_{>1}$	top=VG	% VG	# Cat
people	51.9	2.1	8.6	4.3	49.8	32.3	4533	43.8	2.9	88.5	45.1	20.0	10.9	55
clothing	63.9	1.6	6.4	3.2	70.2	52.6	2192	50.6	2.5	68.3	32.5	38.5	24.6	39
home	66.4	1.5	6.3	3.1	78.5	58.8	6292	50.7	2.7	90.6	42.6	39.3	24.9	89
buildings	66.9	1.5	6.9	3.0	72.6	55.5	967	47.8	2.9	59.9	27.2	27.8	19.2	36
food	71.3	1.3	5.5	2.9	62.9	52.1	1975	47.0	2.5	31.5	15.0	29.3	19.3	92
vehicles	72.0	1.1	4.7	2.4	71.1	60.2	4552	56.5	2.0	63.3	30.0	18.4	17.9	49
animals,plants	91.3	0.4	2.7	1.5	93.8	88.0	4804	67.6	1.5	26.5	12.3	28.1	25.7	89
all	69.7	1.3	5.7	2.9	72.8	58.7	25315	52.8	2.4	58.2	27.8	29.4	20.9	449

Table 1: Agreement in naming measured on the level of instances and on the level of VG categories (i.e. after grouping objects by their VG name)

VG name	top5 MN names	n_{obj}
Canonical VG names with max agreement in MN		
giraffe	giraffe (96.8), animal (1.2), zebra (0.4), camel (0.3), pole (0.1)	915
zebra	zebra (96.3), animal (1.0), giraffe (0.9), horse (0.2), microwave (0.2)	461
cat	cat (94.8), animal (0.9), kitten (0.8), dog (0.4), laptop (0.2)	754
Canonical VG names with min agreement in MN		
booth	booth (19.3), table (12.3), phone booth (9.8), bench (6.7), building (4.4)	11
cabbage	cabbage (21.4), lettuce (17.0), hotdog (11.9), food (10.7), salad (10.4)	9
robe	robe (22.1), shirt (16.8), jacket (13.3), dress (5.7), clothing (3.2)	19
Non-canon. VG names with max agreement in MN		
sedan	car (88.4), wheel (3.1), vehicle (2.3), automobile (1.3), dog (0.8)	11
pony	horse (83.9), pony (9.1), animal (2.9), donkey (1.1), cow (1.1)	8
necktie	tie (81.4), necktie (10.2), shirt (4.6), ties (1.5), jacket (0.5)	11
Non-canon. VG names with min agreement in MN		
shelter	umbrella (9.7), shelter (8.8), roof (8.0), tent (7.1), building (6.8)	10
bath	shower (13.3), elephant (9.9), bird-bath (8.1), water (7.2), trough (7.2)	10
vegetable	food (15.7), broccoli (13.1), sandwich (10.6), salad (9.3), pizza (7.8)	25

Table 2: Qualitative examples for VG names and their most frequent corresponding responses in the MN data set (percentages shown in brackets), canonical means that the VG name is the top name in MN, and non-canonical vice versa