Object naming in the wild: *//gb: add sthg make more concrete//*

Anonymous ACL submission

Abstract

1 Introduction

The real-world objects that we interact with in our every-day life can be categorized into many thousands and maybe millions of categories. And even a single object can be member of many categories, i.e. at different taxonomical levels or in different parts of a taxonomy. For instance, both objects in Figure 1 are at once instances of CAKE, CHEESECAKE, DESSERT, SWEET, PASTRY, FOOD etc.

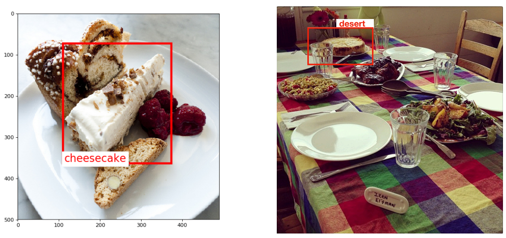

Figure 1: Two objects of the same type of cake, with different names in VisualGenome

Given the abundance of concepts available in language, the act of *naming* a visual object is not just a labeling of visible properties, it amounts to selecting a name from a complex network of concepts and competing lexical alternatives. Hence, research on cognition and language production has relied on object naming as a basic paradigm for investigating the processes that underly formation and organization of concepts in the human mind (Rosch et al., 1976) *//sz: cite more here//*, though mostly using idealized, graphical objects from specific domains (plants, animals) as visual stimuli. Complementary to that, research in computer vision has (successfully) focused on automatically recognizing *real-world* objects in images or videos, but using simplified categorization schemes where each object is assigned a single correct label or name, cf. (Szegedy et al., 2015).

In NLP, to date, research on object naming is relatively scarce despite the fact that there has been a recent explosion of interest in various, and even complex, language & vision tasks ranging from image captioning (Fang et al., 2015; Devlin et al., 2015; Bernardi et al., 2016) to e.g. visual dialogue (Das et al., 2017; De Vries et al., 2017). Massive data collections for applications in language & vision (L&V) are nowadays available and, in principle, these should also constitute an excellent, large-scale test bed for assessing theories such as, e.g. , the claim that objects have a preferred entry-level when being named (Rosch et al., 1976).

The goal of this paper is to extend Visual Genome (Krishna et al., 2016), a well-known, large-scale resource in language & vision research, in a way that it can serve as a broad empirical basis for systematic and linguistically motivated investigations into object naming. We argue that object naming is an interesting, core phenomenon in itself as it occurs in virtually every L&V task, but our approach can also support more systematic analysis of broader tasks, such as e.g. modeling referring expressions.

Even though VisualGenome is one of the most exhaustively annotated resources to date, providing dense object annotations and descriptions in real-world images, it suffers from two important shortcomings if one is interested in linguistic analysis of object naming: First, it only provides a single, manually annotated object description (including a single name) per object which makes it impossible to assess how representative the annotated naming choices are, e.g. whether speakers tend to generally prefer *cheesecake* for the highlighted object in Figure 1. Second, it does not provide consistent taxonomic information on objects and their categories, as names have been automatically linked to WordNet synsets. This makes it difficult to assess how naming depends on the taxonomic properties of the object, e.g. that both objects in Figure 1 are instances of CHEESECAKE, but one is named *cheesecake* and the other one is named *desert*. It is important to note here that these shortcomings exist for basically all large-scale resources currently used in L &V research (see Section 2 below).

In this work, we address these two shortcomings and present a crowdsourcing-based and lightweight experimental set-up for eliciting representative and taxonomically consistent (*//sz: more complete?//*) naming data. We compare our collected data against names annotated in Visual Genome, and calculate various measures assessing agreement, naming preferences etc.

The main idea is to elicit names in (i) in a standard naming task (phase 0) where participants simply give the most straightforward name to an object they can immediately think of,

2 Related Work

Cognition: Concepts and categorization Seminal work on concepts by Rosch suggests that object names typically exhibit a preferred level of specificity called the **entry-level**. This typically corresponds to an intermediate level of specificity, i.e., **basic level** (e.g, *bird*, *car*) (Rosch et al., 1976), as opposed to more generic (i.e., **super-level**; e.g., *animal*, *vehicle*) or specific categories (i.e., **sub-level**; e.g., *sparrow*, *convertible*). However, less prototypical members of basic-level categories tend to be instead identified with sub-level categories (e.g., a PENGUIN is typically called a *penguin* and not a *bird*) (Jolicoeur, 1984). While the traditional notion of entry-level categories suggests that objects tend to be named by a *single* preferred concept, research on pragmatics has found that speakers are flexible in their choice of the level of specificity. Scenarios where multiple objects (of the same category) are present induce a pressure for generating names which uniquely identify the target (Olson, 1970), such that sub-level names can be systematically elicited in these cases (Rohde et al., 2012; Graf et al., 2016).

Vision: Object Recognition State-of-the-art computer vision systems are able to classify images into thousands of different categories (e.g. Szegedy et al. (2015)). These object recognition systems are now widely used in vision & language research. Nevertheless, the way the treat object recognition is conceptually very simple (if not to say, naive): standard object classification schemes are inherently "flat", and treat object labels as mutually exclusive (Deng et al., 2014), ignoring all kinds of linguistic relations between these labels and ignoring the fact that an object can easily be an instance of several categories.*//cs: I would make this statement stronger and argue that* ***object recognition is merely a labeling of objects with human interpretable symbols****, and that a system would probably fail if it had to decide whether an object labeled as, e.g. fig may also be labeled as food.//*

Vision & language: Naming and Referring Ordonez et al. (2016) have studied the problem of deriving appropriate object names, or so-called entry-level categories, from the output of an object recognizer. Their approach focusses on linking abstract object categories in ImageNet to actual words via translation procedures that e.g. involve corpus frequencies. Zarrieß and Schlangen (2017) learn a model of object naming on a corpus of referring expressions paired with objects in real-world images, but focus on combining visual and distributional information and on zero-shot learning. Thus, object naming is an important task for referring expression generation, though most research in this area has focussed on content and attribute selection (Kazemzadeh et al., 2014; Gkatzia et al., 2015; Zarrieß and Schlangen, 2016; Mao et al., 2015).

Existing resources and their shortcomings Moreover, existing resources in L&V hardly provide any consistent taxonomic information on objects and their categories, e.g. object labels are typically quite general as in Flickr30k (Plummer et al., 2015, e.g., PEOPLE, ANIMALS, BODY-PARTS, CLOTHING) or taxonomically heterogeneous as in MS COCO (Lin et al., 2014, e.g., PEOPLE, BASEBALL GLOVE, BIRD).

3 Data Collection

describe the task here

3.1 Visual Genome data

//gb: START to be refined – taken from sivl submission as is//

VisualGenome (Krishna et al., 2016) aims to provide a full set of descriptions of the scenes which images depict in order to spur complete scene understanding. It contains a dense region-based labeling of $108k$ images with textual expression of the attributes and references of objects, their relationships as well as question answer pairs, all linked to WordNet synsets (Fellbaum, 1998, see below).

(1) **Specific categories**: are not available, as object categories and names are not consistently annotated (and even conflated)

(2) **Exhaustive annotations**: are available, which is a huge advantage of this data sets

(3) **Natural names**: are available, though object names might not be fully discriminative (as in referring expressions)

//gb: END to be refined – taken from sivl submission as is//

//cs: START @ GBT//

3.2 Sampling of Instances (Images/Objects)

Since our work connects research on object naming in computer vision and in psycholinguistics/cognitive science, we aimed at collecting a relatively large amount of naturalistic images (*instances*) that depict objects of frequent classes/names in visual genome, which, at the same time, have been frequently/commonly studied in the psycholinguistic literature. We chose the concepts of McRae et al.'s feature norms (REF), which are common objects of different categories (e.g., ANIMALS, FURNITURE) and, as such, have a high overlap with standard datasets of norming studies (REFS). In contrast to the latter, the McRae norms do not contain names of the PERSON category, which we manually added.

(As appropriate: We use image and object interchangeably in the following, since we only selected one target object per image (i.e., each object and image in VG is chosen at most once).)

Collection nodes We defined a set of *collection nodes* which we would then use to collect our object instances from VG.

We based the definition of our set of nodes on the WN (REF) synsets of the McRae concepts (e.g., dog, duck, goose, gull), the nominal WordNet hierarchy, and the frequency distribution of the VG object names' synsets.[1]

First, we selected a set of collection node candidates—synsets which match (e.g., *dog, duck, goose, gull*) or subsume (e.g., *mammal, bird*) the McRae synsets[2]. From these candidates we kept those as collection nodes which had a high frequency of VG object instances of different names. For example, VG instances subsumed by McRae's *dog* were named *beagle, greyhound, puppy, bulldog*, etc., while McRae's *duck, goose*, or *gull* did not have name variants in VG, so we kept *dog* and *bird* as collection nodes.

Goal of this procedure is the collection of instances of selected object classes—our nodes—whose VG names correspond to or subsume (are hypernyms of) a McRae concept, and whose object names differ, that is, of which we can expect that people possess different names for them (e.g., *duck, goose, gull* for *bird*). The collection of such instances using the nodes was then straightforward: We retrieved all VG images depicting an object whose name matches or is subsumed by one of the collection nodes. We did not consider objects with names in plural form, with parts-of-speech other than nouns[3], or that were multi-word expressions/phrases (e.g., *pink bird*). We further only considered objects whose bounding box[4] have an area of $20 - 90\%$ of the whole image area.

Finally, from this set of instances we sampled our final dataset of $31,093$ instances. Sampling proceeded in dependence on the overall size of the individual collection seeds: up to 800 objects per seed: all instances, but at most 500, are collected; more than 800 objects per seed: all instances, but at most $1,000$, are collected. **double-check**

//cs: END @ GBT//

Table **??** gives an overview of the collection nodes, XXX, XXX, grouped into 7 domains. (**Report only dataset after round0, with a note in caption/footnote referring to the checkpoint pruning.**)

Number of images/objects: 25,596
Number of object names: 450

[1]TODO: need to be clear from the general description of VG that the frequ. of instances labeled with the synset of the object name is meant.

[2]Specific synset IDs, e.g., dog.n.01, are omitted for readability.

[3](REF to tagger)

[4]TODO: need to be clear from the general description of VG what is meant.

Number of collection nodes (synsets): 52

3.3 Procedure

describe the crowdsourcing set-up and the task TODO: Footnote: we ran pilot experiments to design our experiment and instructions.

Collection Method

- instructions; put layout in appendix
- each round: HIT of 10 instances, collect 9 annotations for each HIT
- round 0 (with opt-outs) → pruning → rounds 1-3 (no opt-outs)
  pruning: Based on given opt-outs: keep images with no OCCLUSION, at most BBOX is ambiguous twice, at most 17% of names in plural form, most frequent names is of same domain as VG name (gives $25,596$, i.e., discard $5,497$ instances)
- workers could only participate in one round, such as to avoid workers annotating an instance more than once.

Overall XX participants, each annotated between XX and XX instances.

3.4 Data

give an overview of the collected data

4 Analysis

//gb: Note: the structure below is not supposed to be the one in the paper; it was the easiest way for me to plan the analysis a bit//

4.1 Agreement, basic-level and entry-level names

analyse data from Phase 0

Items:

- to what extent do people agree when their task is to give the most straightforward name they can think of to a visual object? (see 4.1.1)
- is the level of agreement the same for all categories? (see 4.1.1)
- how specific are the most familiar names? link names to WordNet, show that WordNet might not be ideal to assess specificity
- assess how representative name annotations in Visual Genome are, when compared to our names (see 4.1.2)

4.1.1 Agreement: Snowgrad's measure

Note: I (Gemma) will use three levels of analysis: ALL (all data lumped together), DOMAIN (Gemma's reorganization of Carina's "supercategories"; see doc 0_object_naming_taboo), COLLECTION NODE (Carina's "synset / collection node").

Plans for analysis (then we see what to put in the paper):

1. compute snowgrad measure and do a:
   - histogram ALL
   - boxplot by DOMAIN
   - dataframe with mean and sd by COLLECTION NODE

   → This will tell us how much agreement there is among subjects about how to name objects in general and within each domain/"subcategory".
2. can we find generalizations about tendencies in agreement? (open: how to go about it)

4.1.2 How representative are VG names?

Compute the most frequent name for each image and see how often that name coincides with the one given by the VG annotator.

4.2 Cases of disagreement

when and why do people give different names to the same object? this will probably happen in phase 0, and even more so in the later round *//sz: this is what I expect//*

//gb: Maybe we can fuse this analysis with the one in the next subsection (taxonomic relations). The way I would put it is, instead of disagreement and taxonomy, "sources of variation"//

- analyse naming disagreement using WordNet, how do names for the same object relate to each other according to WordNet?Overarching question:
- can we identify instances of cross-classification? so objects that are systematically part of several classes (e.g. cake/dessert)

- *//cs: Wrt points 1+2: Show that WordNet, again, is not ideal for retrieving all possible names of an object from a single synset.//*
- we might need to do some manual annotation here and try to carefully describe the phenomena

4.3 Taxonomic relations

can we elicit natural sub-ordinate, super-ordinate concepts?

5 Conclusion

We have presented a systematic, large-scale study on object naming with real-world images and crowdsourced data.

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.* 55(1):409–442. http://dl.acm.org/citation.cfm?id=3013558.3013571.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*. Springer, pages 48–64.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 100–105.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*. IEEE, Boston, MA, USA.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the virtual to the realworld: Referring to objects in real-world spatial scenes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1936–1942. http://aclweb.org/anthology/D15-1224.

Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? level of abstraction in nominal referring expressions. In *Proceedings of the 38th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Pierre Jolicoeur. 1984. Pictures and names: Making the connection. *Cognitive psychology* 16:243–275.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Doha, Qatar, pages 787–798.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C.Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision ECCV 2014*, Springer International Publishing, volume 8693, pages 740–755.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *ArXiv / CoRR* abs/1511.02283. http://arxiv.org/abs/1511.02283.

David R Olson. 1970. Language and thought: Aspects of a cognitive theory of semantics. *Psychological review* 77(4):257.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2016. Learning to name objects. *Commun. ACM* 59(3):108–115.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer

Image-to-Sentence Models. *CoRR* abs/1505.04870. http://arxiv.org/abs/1505.04870.

Hannah Rohde, Scott Seyfarth, Brady Clark, Gerhard Jäger, and Stefan Kaufmann. 2012. Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue*. pages 107–116.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology* 8(3):382–439.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*. Boston, MA, USA.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 610–620. http://www.aclweb.org/anthology/P16-1058.

Sina Zarrieß and David Schlangen. 2017. Obtaining referential word meanings from visual and distributional information: Experiments on object naming. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 243–254. http://aclweb.org/anthology/P17-1023.