Object naming is context dependent - A case study on testing linguistic hypotheses on real-world image corpora

Anonymous ECCV submission

Paper ID ***

Abstract. The abstract should summarize the contents of the paper. LNCS guidelines indicate it should be at least 70 and at most 150 words. It should be set in 9-point font size and should be inset 1.0 cm from the right and left margins. . . .

Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

Understanding and modeling the way humans converse about their environment using natural language has been a long-standing goal of research in various fields related to artificial intelligence and linguistics. With recent developments in computer vision and a range of massive data collections in particular, there has been a veritable explosion of interest in language & vision tasks, ranging from image captioning [1–5], referring expression resolution and generation [6–9], to multimodal summarization or visual dialogue [10, 11]. In principle, the underlying data collections here should not only spur computational, application-oriented research aimed at implementing systems for very specific tasks – they should also constitute extremely valuable resources for research aimed at deriving linguistic generalizations about various phenomena related to language grounding, reference and situated interaction which, for a long time, have been investigated mostly in very controlled and small-domain experimental settings, cf. [**?**,**?**,12, **?**,**?**] for some examples of traditional data collections related to reference and grounding. In turn, these linguistic generalizations could inform computational modeling, architecture design and future data collections. However, so far, studies that have tested linguistic hypotheses on large-scale vision & language resources have been relatively rare.

In this paper, we take a look at object naming, a core phenomenon that occurs in virtually every language & vision task and is, at the same time, subject of ongoing research in language grounding and pragmatics. Our starting point is a particular linguistic hypothesis related to object naming - namely that the choice of a name for an object is dependent on other objects in its visual context - which has been tested recently in a classical experimental setting [13]. We discuss how this hypothesis could be tested based on recent data sets that pair images and object descriptions, referring expressions or captions (all containing object names) and define a set of requirements for obtaining a theoretically informed model for object naming. In sum, this discussion will show that specific requirements are met by particular resources, but none of the available corpora consistently satisfies all of the requirements. We believe that this is a perfect showcase illustrating the challenges for linguistically motivated research in language and vision, and we derive a proposal for obtaining more consistently designed and annotated data collections for object naming.

2 Naming Objects (in Context)

The act of naming an object amounts to that of picking out a nominal to be employed to refer to it (e.g., “the *dog*”, “the white *dog* to the left”). Since an object is simultaneously a member of multiple categories (e.g., a young beagle is at once a dog, a beagle, an animal, a puppy etc.), all the various names that lexicalize these constitute a valid alternative, meaning that the same object can be named with more or less **specific names** [14, 15]. 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*) [16], 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*) [17]. This out-of-context preference towards a certain taxonomic level is often referred to as **lexical availability**.

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 with respect to the chose level of specificity. Scenarios where multiple objects (of the same category) are present induce a pressure for generating names which uniquely identify the target [18], such that sub-level names can be systematically elicited in these cases [19] [13]. For example, in presence of more than one dog, the name *dog* is ambiguous and a sub-level category (e.g., *rottweiler*, *beagle*) is more informative and potentially preferred by speakers, though additional factors such as cost (which is typically approximated by frequency) or saliency also come into play [13] [20].

So far, research in computer vision, and vision & language, has mostly worked around the fact that objects can be categorized and named at different levels of specificity. State-of-the-art object recognizers are typically trained on a flat set of categories taken from ImageNet (REF!!), though see work by Deng et al. on trading off object recognition accuracy and level of specificity [21]. [22] present one of the few explicit studies on naming in computer vision and operationalize the task as translating between leaf nodes in the ImageNet hierarchy and entry-level concepts, adopting the traditional view that there is a single preferred name for a given object. Thus, establishing whether object naming is context dependent in realistic, large-scale data sets would not only be of interest to theoretical

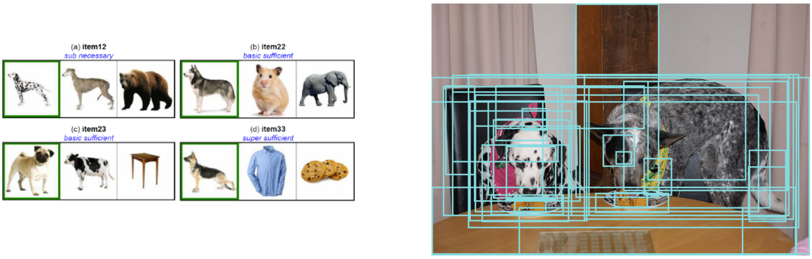

Fig. 1. Experimental and real-world visual scenes showing dogs (GET MORE IMAGES OF DOGS FROM VG???)

work on concepts and pragmatics, it would also be of great importance for the design of models in computer vision, object recognition and vision & language.

3 Requirements

In this Section we discuss requirements for scaling experimental studies on object names as in [13] to real-world images. The difference between these two approaches is illustrated in Figure 1, showing [13]'s carefully controlled experimental conditions with isolated objects arranged in a collage, and a real-world scene where multiple objects occur in a natural context. Thus, in contrast to [13] we do not aim for assembling controlled and, to some extent, artificial scenes, but we aim for studying object naming as prompted by real-world images. Therefore, we not only need access to names naturally generated by speakers, but also, we need to be able to quantify those factors whose interaction with naming we want to analyze.

Given a natural scene with multiple objects and a target object that a speaker referred to, we need to be able to answer the following questions:

R0 **Lexical choice**:
Which name did the speaker use to refer to the target object?

R1 **Lexical availability**:
Which names are available for the target objects? For instance, if we want to check whether a convertible is more often referred to with a basic-level category (*car*) or a more specific one (*convertible*) we need to know that a given object is a *convertible*, and that *convertible* is a sub-level concept of *car* (and vehicle, etc.). Additionally, in [13], the authors manually group object names used by speakers in terms of three levels (sub-level, basic-level, super-level). Thus, ideally, names would be grouped according to their taxonomic relations.

R2 **Contextual informativeness**:
Which names are available for the other objects in the scene (i.e. distractors)? For instance, when the target is a convertible, we need to know that there are multiple convertible, cars, etc. to judge whether these concepts are ambiguous.

4 Resources: What Do We Have?

5 Proposal

References

1. Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., Zweig, G.: From captions to visual concepts and back. In: Proceedings of CVPR, Boston, MA, USA, IEEE (June 2015)
2. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: 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), Beijing, China, Association for Computational Linguistics (July 2015) 100–105
3. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2422–2431
4. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition. (2015)
5. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. J. Artif. Int. Res. **55**(1) (January 2016) 409–442
6. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: 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 (2014) 787–798
7. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. CoRR **abs/1511.02283** (2015)
8. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L. In: Modeling Context in Referring Expressions. Springer International Publishing, Cham (2016) 69–85
9. Schlangen, D., Zarriess, S., Kennington, C.: Resolving references to objects in photographs using the words-as-classifiers model. In: Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016). (2016)
10. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2017)
11. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proc. of CVPR. (2017)
12. Krahmer, E., Van Deemter, K.: Computational generation of referring expressions: A survey. Computational Linguistics **38**(1) (2012) 173–218
13. Graf, C., Degen, J., Hawkins, R.X., Goodman, N.D.: 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 (2016)
14. Brown, R.: How shall a thing be called? Psychological review **65**(1) (1958) 14
15. Murphy, G.: The big book of concepts. MIT press (2004)
16. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. Cognitive psychology **8**(3) (1976) 382–439
17. Jolicoeur, P.: Pictures and names: Making the connection. Cognitive psychology **16** (1984) 243–275

18. Olson, D.R.: Language and thought: Aspects of a cognitive theory of semantics. Psychological review **77**(4) (1970) 257
19. Rohde, H., Seyfarth, S., Clark, B., Jäger, G., Kaufmann, S.: Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue. (2012) 107–116
20. Clark, H.H., Schreuder, R., Buttrick, S.: Common ground at the understanding of demonstrative reference. Journal of verbal learning and verbal behavior **22**(2) (1983) 245–258
21. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3450–3457
22. Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: Learning to name objects. Commun. ACM **59**(3) (February 2016) 108–115
23. Fellbaum, C.: WordNet. Wiley Online Library (1998)