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 [1,2,3,4,5], referring expression resolution and generation [6,7,8,9], to multi-modal 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 that can be determined independentl of the visual context.

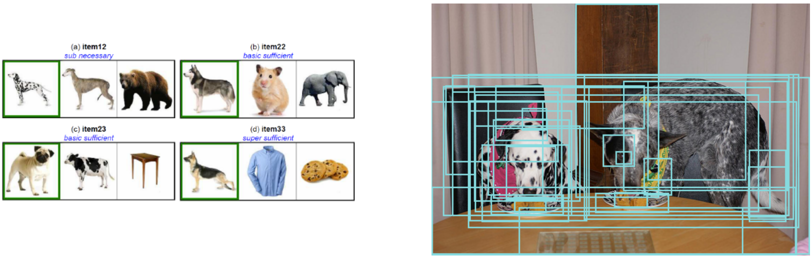

Fig. 1. Experimental and real-world visual scenes showing dogs (GET MORE IMAGES OF DOGS FROM VG???)

3 Requirements

In this Section we discuss the requirements for being able to study the context dependence of object names to real-world images, inspired by [13]'s study. Figure 1 shows [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 from Visual Genome [23]. Thus, in contrast to [13] we do not aim for assembling controlled and, to some extent, artificial scenes, meaning that we need to be able to quantify contextual factors for naming in a real-world image corpus. Given a real-world scene with multiple objects and a target object that has to be named, we need access to:

(1) the **specific category** of the object that the speaker referred to
(2) **exhaustive annotations** for all the other objects in the scene (i.e. distractors)
(3) the **natural name** chosen by a speaker for describing or referring to an object

For instance, if we want to check whether a specific object like convertible is more often referred to with a basic-level category (e.g., *car*) or a more specific one (e.g., *convertible,limousine*), all three requirements need to be fulfilled: we need to know whether the object is a CONVERTIBLE or a LIMOUSINE, and likewise for its distractors, and we need to know wether a speaker called the object that *convertible,car, etc.*

4 Resources: What Do We Have?

4.1 ImageNet and ILSVRC

ImageNet [24] is one of the biggest publicly available image databases that follows a consistent taxonomy: its hierarchical organization is based on WordNet [25], such that many synsets from WordNet can be queried for visual instances. Vice versa, given a leaf node in the ImageNet hierarchy, all its specific and more general names can be retrieved by following the links between synsets in WordNet (as is actually done by [22]). Many state-of-the-art object recognizers are trained on a subsample of ImageNet. However, the database is not directly usable for our purposes as each image only depicts a single object (i.e. an instance of the synset), such that no (or very few) distractors objects.

4.2 RefCOCO and RefCOCO+ [8]

Both datasets are extensions of ReferIt [6], a large-scale collection of referring expressions (RE) for natural objects in real-world images, and are built on top of the MS COCO image collection [26], a dataset of images of natural scenes of 91 common object categories (e.g., DOG, PIZZA, CHAIR). The REs were collected via crowdsourcing in a two-player game, which was designed to obtain REs uniquely referring to the target objects in an image. Specifically, a director and a matcher are presented with an image, and the director produces a RE for an outlined target object in the image. The matcher must click on the object he thinks the RE refers to. If the matcher's prediction is correct, the RE is considered valid.

Shortcomings REs in RefCOCO and RefCOCO+ were collected under the constraints that (i) all images contain at least two objects of the same category (80 COCO categories), which prompts the players to avoid the mere object category as RE, and (ii) in RefCOCO+ the players must not use location words, urging them to refer to the appearance of objects. Another critical property of the data is that, (iii), not all objects in an image were annotated with REs, may it due to the frequency constraint (i), or due to the object not being part of the 80 COCO categories. Finally, the 80 COCO categories tend to be entry-level categories and are not linked to the ImageNet taxonomy. *//cs: TODO: EX / ANALYSIS//*

(1) **Specific categories**: are not available, as we only have access to basic-level categories and cannot retrieve the most specific category or name for each object
(2) **Exhaustive annotations**: are not available, as not all objects were annotated with REs and corresponding categories
(3) **Natural names**: are available, though it is unclear how the additional constraints in RefCoco+ impact on the naturalness of object naming

4.3 Flickr30k Entities

The Flickr30k Entities dataset [27][1] augments Flickr30k, a dataset of 30k images and five sentence-level captions for each of the images, with region-level annotations. Specifically, mentions of the same entities across the five captions of an

[1] Available at `web.engr.illinois.edu/~bplumme2/Flickr30kEntities`

image are linked to the bounding boxes of the objects they refer to. The dataset was designed to advance image description generation and phrase localization in particular (e.g., [28,29,30]).

Shortcomings By design, Flickr30k Entities can be used to study the way people refer to individual entities in an image depending on the situation the speakers describe and, in contrast to RefCOCO (+), the production of entity mentions did not underlie any constraints. On the other hand, Flickr30k Entities is less suited for referring expression generation since mentions in isolation of their linguistic context may not uniquely identify the referred object.

(1) **Specific categories**: are not available, object categories tend to be even less specific than those of COCO (e.g., PEOPLE, ANIMALS, BODYPARTS, CLOTHING), or are abstract (OTHER, SCENE)
(2) **Exhaustive annotations**: are not available
(3) **Natural names**: are available, though object names might not be fully discriminative (as in referring expressions)

4.4 Visual Genome

VisualGenome [23] 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 [25, see below].

Shortcomings

(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)

5 Possible Solutions

We now discuss solutions to the shortcomings discussed above.

5.1 Object detectors

As discussed in Section 4, none of the corpora that contain natural naming data for objects in context provide consistent annotations of object categories at a sufficient level of specificity. An alternative would be to apply object detectors or image classifiers trained to predict the most specific category of the full inventory of objects which the dataset covers. However, pre-trained models only exist for a subset of the datasets' objects. *//cs: TODO: ADD CMP WITH ILSVRC//* For the training of a model using, e.g., ImageNet [24], on the other hand, the set of most-specific categories covered by the data needs to be provided or collected from humans.

5.2 Post-annotating Captions

Regarding RefCOCO/+, additional REs could be collected from the COCO Captions dataset. Since annotators could naturally refer to depicted objects in their descriptions (i.e., choose the entry-level as the most preferred reference whenever possible), we may infer the entry-level of the objects from them through maximum likelihood estimation. In contrast to Flickr30k Entities, though, the data does not contain region-phrase associations, such that natural language phrases first needed to be aligned with the image regions they refer to. This task of *language grounding*, which has been an active research topic in vision & language (e.g., [31,32,28]), is beyond the focus of our object naming study.

5.3 Collect naming data with controlled specificity

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. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. (2016)
24. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. (2009)
25. Fellbaum, C.: WordNet. Wiley Online Library (1998)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., Zitnick, C.: Microsoft coco: Common objects in context. In: Computer Vision ECCV 2014. Volume 8693. Springer International Publishing (2014) 740–755
27. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. CoRR **abs/1505.04870** (2015)
28. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV 2016). (2016)
29. Plummer, B.A., Mallya, A., Cervantes, C.M., Hockenmaier, J., Lazebnik, S.: Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In: Proceedings of the International Conference on Computer Vision (ICCV 2017). (2017) 1946–1955
30. Yeh, R.A., Do, M.N., Schwing, A.G.: Unsupervised Textual Grounding: Linking Words to Image Concepts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). (2018)
31. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are You Talking About? Text-to-Image Coreference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). (2014)
32. Karpathy, A., Fei-Fei, L.: Deep Visual-semantic Alignments for Generating Image Descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). (2015) 3128–3137