How do we get from object recognition to object naming?

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

Real-world objects are members of many categories. and speakers can typically between choose between different, more or less specific names when referring to a particular visual entity. For instance, the entity surrounded by the green box in Figure 1 is an instance of the categories *female child, child, female, person, organism*, etc. and can be referred to with names such as e.g. *girl, kid, cutie, daughter, person, human*. But even though almost every task in language & vision involves the prediction of object names (e.g. image captioning, referring expression generation, visual dialogue), hardly any research has explicitly looked at object naming, i.e. determining the actual word/linguistic concept that speakers would use to refer to an object (in a particular context), but see [1, 2]. Even research in pragmatics, that traditionally deals with reference and referring expression production, has mostly focussed on attributes, rather than object names. Recently, however, [3] have shown that object naming preferences are subject to contextual constraints and pragmatic factors: in a typical reference game set-up with images of target objects surrounded by distractor objects, speakers have been found to flexibly adjust object names depending on the context. For instance, a *dalmatian* would be called *dalmatian* in the context of other dogs or simply *dog* when none of the distractors is also a dog. This extends previous traditional work on concepts suggesting that the typicality of a referent determines its entry-level category (and consequently, its name) [4].

Most research on computer vision and object recognition has somehow worked around the fact that objects can be instances of multiple, hierarchically embedded categories. State-of-the-art recognition models are mostly trained as classifiers on a flat of 1000 categories in ImageNet, reducing the ImageNet ontology to an inherently flat annotation scheme and treating object labels as mutually exclusive.

On the vision side, however, there is huge body of research on object recognition, i.e. labeling visual objects according to a set of categories, cf. [5–7].

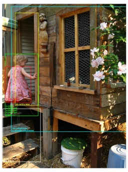

Fig. 1. INCLUDE MSCOCO examples

2 Background

3 Data

3.1 Corpora

Referring Expressions We analyze the RefCOCO and RefCOCO+ datasets which contain referring expressions to objects in MSCOCO [8] images. These data collections were performed via crowdsourcing with the ReferIt Game [9] where two players were paired and a director needed to refer to a predetermined object to a matcher, who then selected it. RefCOCO and RefCOCO+ contain 3 referring expressions on average per object, and overall 150K expressions for 50K objects. The two datasets have been collected for an (almost) identical set of objects, but in RefCOCO+, players were asked not to use location words (*on the left*, etc.). See [10] for more details.

Image Captions TODO.

3.2 Preprocessing

We parse referring expressions and captions with the Stanford Dependency Parser. We extract heads/object names as follows: TODO.

4 Names: Levels of specificity

In this Section, we investigate whether variability of reference level can be observed in existing data sets for language & vision.

4.1 Using WordNet

[3] investigate object naming with respect to reference level. They distinguish and manually annotate 3 levels: (i) sub-level (*dalmatian*), (ii) basic-level (*dog*), (iii) super-level (*animal*).

For large-scale studies of object naming, we need to be able to automatically define the level of specificity of a name, given an ontology. In this Section, we investigate whether WordNet is appropriate for defining reference level. We hypothesize that the distance of a name's synset to the root node (entity) relates to its specificity.

Specifcity We calculate this distance as follows: we lookup all synsets of a word and retrieve the respective paths to the root node in WordNet. For each word, we use the minimal path length as distance to the root node.

Table 1 shows the levels of specificity we observe for object names in the RefCoco data set. We observe distances to the root between 2 and 17, meaning that there is a much more fine-grained distinction of levels as the three-way classification adopted by [3].

Unfortunately, the levels of specificity predicted by WordNet do not seem to reflect linguistic intuitions, here are some problematic examples from Table 1:

- elephant (10) is more specific than panda (14)? horse is less specific than elephant (10)?

specificity	rel.freq.	top 5 names
-1	0.071697	NONE,brocolli,zeb,broc,girafe
2	0.003898	thing,things
3	0.001182	object,group,set,substance,objects
4	0.140633	man,person,piece,head,part
5	0.100739	player,glass,baby,front,corner
6	0.208590	woman,girl,kid,boy,bowl
7	0.238708	guy,right,chair,lady,bear
8	0.110613	horse,bus,cow,pizza,batter
9	0.097390	shirt,car,bike,donut,catcher
10	0.048368	elephant,couch,truck,vase,suitcase
11	0.008828	motorcycle,clock,mom,dad,scissors
12	0.002822	oven,airplane,suv,taxi,refrigerator
13	0.005253	laptop,fridge,canoe,orioles,pigeon
14	0.000414	panda,freezer,penguin,rooster,rhino
15	0.030870	zebra,giraffe,zebras,giraffes,deer
16	0.000083	bison,mooses,orang,elks,sambar
17	0.000143	ox,cattle,gnu,mustang,orca

Table 1. Levels of specificity for naming choices in RefCOCO: for each level, relative frequency and 5 most frequent names are shown

References

1. 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
2. Zarrieß, S., Schlangen, D.: 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), Vancouver, Canada, Association for Computational Linguistics (July 2017) 243–254
3. 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)
4. Rosch, E.: Principles of Categorization. In Rosch, E., Lloyd, B.B., eds.: Cognition and Categorization. Lawrence Erlbaum, Hillsdale, N.J., USA (1978) 27—-48
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
6. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs. In: European Conference on Computer Vision, Springer (2014) 48–64
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR 2015, Boston, MA, USA (June 2015)
8. 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
9. 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
10. 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