An object deserves more than a single name

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 different 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, the object in Figure 1 is at once a CAKE, CHEESECAKE, DESSERT, SWEET, PASTRY, FOOD etc.

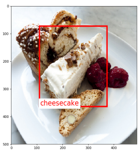

Figure 1: A real-world object in VisualGenome

Given the abundance of concepts available in language, the act of *naming* a visual object is less trivial than it might seem intuitively: in order to describe or refer to an object, a name has to be selected from a complex network of concepts and from many 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). In this paper, we argue that a lot of research in this area would benefit from a deeper and more systematic understanding of the semantic and taxonomic processes in object naming, which is a core phenomenon in virtually every vision & language task. At the same time, we argue that existing resources in L&V constitute an excellent, large-scale test bed for assessing traditional claims and theories such as, e.g. the existence of so-called entry-level categories (Rosch et al., 1976).

The goal of this paper is to extend some existing, large-scale resources available for language & vision research in a way that they can serve as a solid and broad empirical basis for investigations into object naming. Thus, while large amounts images from different domains paired with verbal descriptions are generally available, none of the existing corpora records more than a few annotations per object. Even VisualGenome (Krishna et al., 2016), one of the most exhaustively annotated resources to date, only provides a single, manually annotated object description (including a single name) per object. 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, BODYPARTS, CLOTHING) or taxonomically heterogeneous as in MS COCO (Lin et al., 2014) (e.g., PEOPLE, BASEBALL GLOVE, BIRD). These two shortcomings - few names per object and inconsistent labeling of categories - make it difficult to get reliable and stable predictions as to how speakers will likely name a given object.*//sz: need to make this point more clear//*

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.

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 summarize SIVL paper here

3 Data Collection

describe the YouNameIt task here

3.1 Materials

describe sampling of images, category selection

3.2 Procedure

describe the crowdsourcing set-up and the task

3.3 Data

give an overview of the collected data

4 Analysis

4.1 Agreement, basic-level and entry-level names

analyse data mainly from Phase 0

- to what extent do people agree when their task is to give the most straightforward name they can think of to a visual object?
- is the level of agreement the same for all categories?
- how specific are the most familiar names? link names to WordNet, show that WordNet might not be ideal to assess specificity
- how does agreement evolve in the later rounds of the game? (when people have to avoid taboo names), does agreement increase as the set of names becomes more narrow, or does agreement decrease as people do creative, clever, unexpected things?

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//*

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

5.1 Model 1

Train a simple naming model (classifiers) on original VisualGenome names. Test it on our data. How well does the model predict the most familiar name and the set of available names?

5.2 Model 2

Train a simple naming model (classifiers) on our data, maybe combined with VisualGenome data. How well does it work?

5.3 Model 3

can we model taxonomic/conceptual knowledge more directly? induce a taxonomy? or a multimodal space for objects + names?

6 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.

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 Dollár, 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.