Do resources in Language & Vision favour the study of linguistic variation?

A survey and a new collection of Object Naming data

Author1, Author2, Author3
Affiliation1, Affiliation2, Affiliation3
Address1, Address2, Address3
author1@xxx.yy, author2@zzz.edu, author3@hhh.com
{author1, author5, author9}@abc.org

Abstract

blabla

Keywords: keyword1, keyword2, keyword3

1. Introduction

Generally, research in Language & Vision (L&V) is interested in modeling how speakers *naturally* name, refer to or talk about visual objects and scenes, in contrast to predicting abstract object labels as e.g. in Computer Vision. This typically entails that data collections and models need to account for linguistic variation, as there can hardly ever be a single ground-truth utterance when describing or referring to visual entities. An indeed, variation has been accounted for in the modeling and evaluation of certain L&V tasks like image captioning (Vedantam et al., 2015; Bernardi et al., 2016; Dai et al., 2017).

In principle, the massive data collections now available in L&V 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. (Anderson et al., 1991; Fernández and Schlangen, 2007; Krahmer and Van Deemter, 2012; Takenobu et al., 2012; Zarrieß et al., 2016) 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 L&V resources have been relatively rare.

In this paper, we take a look at object naming, a core phenomenon that occurs in virtually every L&V task and is, at the same time, subject of ongoing research in language grounding and pragmatics. We take stock of existing data sets that provide names for objects in real-world images. We contribute a new dataset, ManyNames, that contains 36 crowd-sourced names for 25K instances from VG.

2. Background

2.1. Object Naming as a Linguistic Phenomenon

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** (Brown, 1958; Murphy, 2004). 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). 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 or saliency also come into play (Graf et al., 2016)(Clark et al., 1983).

2.2. Modeling Object Naming

Though names are prominent in referring expressions, investigated a lot in natural language generation (Dale and Reiter, 1995), this area has focused mostly on the selection of attributes (Krahmer and Van Deemter, 2012). Ordonez et al. (2016) takes up the notion of entry-level categories (Rosch et al., 1976) and transfers an object's predicted fine-grained label to its name using text corpus statistics. Zarrieß and Schlangen (2017) learn a naming model on referring expressions and real-world images, but focus on combining visual and distributional information. Recent experimental work on reference found that the specificity of a name is dependent on the taxonomic relatedness of other objects in context

2.3. Relevant Resources

justify the following selection of resources....

3. Survey: Object Naming in L&V resources

3.1. Visual Genome

VG (Krishna et al., 2016) is one of the most densely and richly annotated resources currently available in L&V. In the following, we will focus on describing aspects immediately relevant to object naming only, while many other annotations are available as well (e.g. questions, paragraphs, etc.)

Collection and annotation procedure VG aims to provide a full set of descriptions of the scenes which images depict in order to spur complete scene understanding. The data collection followed a complex procedure, involving many different rounds of annotation. The first round of the procedure, and the basic backbone for the further rounds, is a collection of region-based descriptions: workers were asked to describe regions in the image and draw boxes around the corresponding area in the image. In this stage, workers were encouraged to annotate In a second independent round (involving new workers), annotators were then asked to process the region descriptions by (i) marking the object names contained in the region description, and (ii) drawing a tight box around the corresponding region. As different region descriptions would potentially mention the same objects, each worker was shown a list of previously marked objects and encouraged to select on existing object rather than annotating a new one.

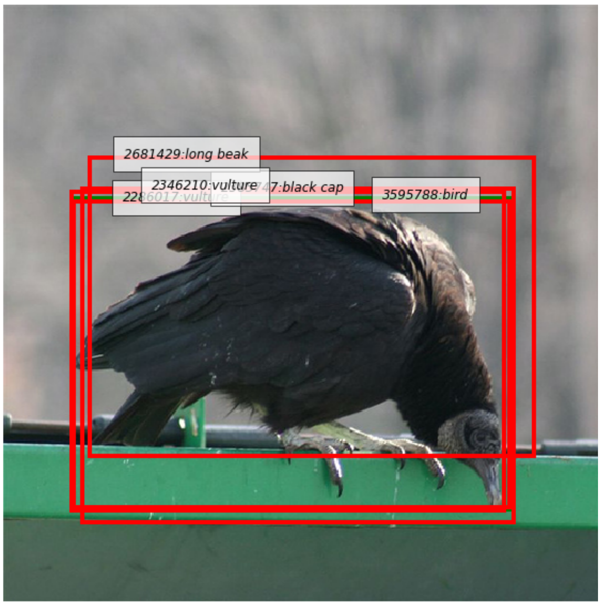
Example Figure 1 shows an example image from VisualGenome, and some of its object annotations. This illustrates that there is only a partial linking of objects that are mentioned across different region descriptions, i.e. the identity of objects cannot be established based on the annotation. for a given object and its bounding box, there might be different region descriptions and names associated with it.

Discussion

- advantages: exhaustive annotations of all/most objects in the image, variable region descriptions and possibly object names
- disadvantages: object linking is partial due to bottom-up annotation procedure

3.2. RefCOCO and RefCOCO+

Both datasets use the ReferIt (Kazemzadeh et al., 2014) game for collecting referring expressions (RE) for natural objects in real-world images, and are built on top of the MS COCO (Lin et al., 2014), 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 reference game designed to obtain REs uniquely referring to the target object. 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. REs in 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.



object id	linked region descriptions
3595788	the **bird** is black in color, nose of the **bird**, a **bird** relaxing in stand, small white beak of **bird**, large black talon of **bird**, a **bird** on a green pole, a green bar under **bird**, black **bird** on green rail, small black eye of **bird**
2286017	large black **vulture** on fence, a vulture on bar
2385747	small white beak of **bird**
2681429	a semi **long beak**
2346210	a black and gray **vulture**

Figure 1: Bounding boxes, names and region descriptions for an object in VisualGenome

(1) **Specific categories**: not available, the 80 COCO categories tend to be entry-level categories and are not linked to the ImageNet taxonomy (e.g., BIRD, PERSON, CAR, BUS)

(2) **Exhaustive annotations**: not available, as not all objects were annotated with REs and corresponding categories

(3) **Natural names**: available, though it is unclear how the additional constraints in RefCoco+ impact on the naturalness of object naming

Analysis We parse REs in RefCOCO with the Stanford Dependency Parser and extract the nominal heads. We map

	RefCoco	Flickr30k	VG	VGmn	MN
# objects	50.000	243.801	3.781.232	25.223	25.315
naming vocab size	5.004	10.423	105.441	1.061	7.970
av. annotations/object	2.84	2.30	1.69	7.24	35.30
% objects with n types > 1	0.68	0.29	0.02	0.05	0.93
av. types/object	1.88	1.38	1.02	1.08	5.70

Table 1: Overview statistics for different data sets containing object naming data

these names to their most frequent sense/synset in WordNet. We hypothesize that the distance of a name's synset to the root node (ENTITY) relates to its specificity. We estimate this distance as the minimal path length of all synsets of a word to the root node. Table 3.2. shows the estimated levels of specificity 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 than the three-way classification adopted in (Graf et al., 2016). Unfortunately, the levels of specificity predicted by WordNet do not seem to reflect linguistic intuitions, e.g. *elephant* is predicted to be more specific than *panda*. At the same time, this overview clearly suggests that object names in RefCOCO do not only comprise entry-level categories, but also very general (*thing*) and very specific names (*ox*).

3.3. Flickr30k Entities

The Flickr30k Entities dataset (Plummer et al., 2015)[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 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., (Rohrbach et al., 2016; Plummer et al., 2017; Yeh et al., 2018)).

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, it 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 REs; e.g., both animals in the right-most image in Fig. **??** are named *dog*)

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

4. Conclusion

Your submission of a finalised contribution for inclusion in the LREC proceedings automatically assigns the above-mentioned copyright to ELRA.

5. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

6. Bibliographical References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Brown, R. (1958). How shall a thing be called? *Psychological review*, 65(1):14.

Clark, H. H., Schreuder, R., and Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258.

Dai, B., Fidler, S., Urtasun, R., and Lin, D. (2017). Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979.

Dale, R. and Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Fernández, R. and Schlangen, D. (2007). Referring under restricted interactivity conditions. In Simon Keizer, et al., editors, *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139, Antwerp, Belgium, September.

Graf, C., Degen, J., Hawkins, R. X., and Goodman, N. D. (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.

Jolicoeur, P. (1984). Pictures and names: Making the connection. *Cognitive psychology*, 16:243–275.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. L. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Con-*

spec.	rel.freq.	top 5 names	spec.	rel.freq.	top 5 names
2	< 0.01	thing,things	10	0.05	elephant,couch,truck,vase,suitcase
3	< 0.01	object,group,set,substance,objects	11	< 0.01	motorcycle,clock,mom,dad,scissors
4	0.14	man,person,piece,head,part	12	< 0.01	oven,airplane,suv,taxi,refrigerator
5	0.10	player,glass,baby,front,corner	13	< 0.01	laptop,fridge,canoe,orioles,pigeon
6	0.21	woman,girl,kid,boy,bowl	14	< 0.01	panda,freezer,penguin,rooster,rhino
7	0.25	guy,right,chair,lady,bear	15	0.03	zebra,giraffe,zebras,giraffes,deer
8	0.11	horse,bus,cow,pizza,batter	16	< 0.01	bison,mooses,orang,elks,sambar
9	0.09	shirt,car,bike,donut,catcher	17	< 0.01	ox,cattle,gnu,mustang,orca

Table 2: Levels of specificity for naming choices in RefCOCO: for each level (distance between name and WordNet root), relative frequency and 5 most frequent names are shown

ference on Empirical Methods in Natural Language Processing (EMNLP 2014), pages 787–798, Doha, Qatar.

Krahmer, E. and Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2016). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., DollÃ¡r, P., and Zitnick, C. (2014). Microsoft coco: Common objects in context. In *Computer Vision - ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Murphy, G. (2004). *The big book of concepts*. MIT press.

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

Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A. C., and Berg, T. L. (2016). Learning to Name Objects. *Commun. ACM*, 59(3):108–115, February.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *CoRR*, abs/1505.04870.

Plummer, B. A., Mallya, A., Cervantes, C. M., Hockenmaier, J., and Lazebnik, S. (2017). Phrase Localization and Visual Relationship Detection with Comprehensive Image-Language Cues. In *Proceedings of the International Conference on Computer Vision (ICCV 2017)*, pages 1946–1955.

Rohde, H., Seyfarth, S., Clark, B., Jäger, G., and Kaufmann, S. (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.

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., and Schiele, B. (2016). Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV 2016)*.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

Takenobu, T., Ryu, I., Asuka, T., and Naoko, K. (2012). The rex corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yeh, R. A., Do, M. N., and Schwing, A. G. (2018). Unsupervised Textual Grounding: Linking Words to Image Concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*.

Zarrieß, S. and Schlangen, D. (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)*, pages 243–254, Vancouver, Canada, July. Association for Computational Linguistics.

Zarrieß, S., Hough, J., Kennington, C., Manuvinakurike, R., DeVault, D., Fernandez, R., and Schlangen, D. (2016). Pentoref: A corpus of spoken references in task-oriented dialogues. In *10th edition of the Language Resources and Evaluation Conference*.