Informativity in Image Captions vs. Referring Expressions

Elizabeth Coppock, Danielle Dionne, Nathanial Graham, Elias Ganem,
Shijie Zhao, Shawn Lin, Wenxing Liu and **Derry Wijaya**
Boston University, MA
ecoppock@bu.edu

1 Introduction

At the intersection between computer vision and natural language processing, there has been recent progress on two natural language generation tasks: *Dense Image Captioning* and *Referring Expression Generation* for objects in complex scenes (Farhadi et al., 2010; Karpathy and Fei-Fei, 2014; Vinyals et al., 2014; Krishna et al., 2017; Mao et al., 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018, 2019). The former aims to provide a caption for a specified object in a complex scene for the benefit of an interlocutor who may not be able to see it, and may form part of a larger Visual Question Answering (VQA) system (Antol et al., 2015; Goyal et al., 2017; Zhang et al., 2016). The latter aims to produce a referring expression that will serve to identify a given object in a scene that the interlocutor can see. The two tasks are designed for different assumptions about the common ground between the interlocutors, and serve very different purposes, although they both associate a linguistic description with an object in a complex scene. Despite these fundamental differences, the distinction between these two tasks is sometimes overlooked (Mao et al., 2016; Cohn-Gordon et al., 2018, 2019). Here, we undertake a side-by-side comparison between image captioning and reference game human datasets and show that they differ systematically with respect to informativity. We hope that an understanding of the systematic differences among these human datasets will ultimately allow them to be leveraged more effectively in the associated engineering tasks.

2 Background and Predictions

As the purpose of using a referring expression is to distinguish one referent from another, without being overly wordy, a naive expectation would be that referring expressions should contain as much information as is necessary to do that, and no more. In other words, descriptive modifiers are expected to be included only if they are *informative* in the sense of helping to narrow down on the set of potential referents. This kind of behavior is predicted by the Rational Speech Act (RSA) framework (Frank and Goodman, 2012): Speakers optimize their choice of expression through a tradeoff between accuracy and cost, and listeners use a Bayesian reasoning process to identify a speaker's referent.

In work on *Referring Expression Generation* (REG; see Krahmer and Van Deemter 2012), RSA has not been viewed entirely without skepticism. Gatt et al. (2013) compare RSA to a Probabilistic Referential Overspecification model (PRO). They conclude that RSA is insufficient because it fails to consider overspecification and preference rankings when generating referring expressions. Baumann et al. (2014) conduct production and interpretation studies that question the assumption that speakers aim to minimize production costs. Their findings suggest that speakers may favor overspecification not only to help the listener, but to avoid the additional cognitive effort.

Amendments to RSA have been proposed in order to account for overinforativity. Degen et al. (2019) do so by adjusting the deterministic semantics that exists in the basic framework to continuous (fuzzy) semantics. Cohn-Gordon et al. (2018) leverage the captions from the Visual Genome corpus (Krishna et al., 2017) in order to define a semantics for an RSA-based referring expression generation system. The incremental nature of their system provides an alternative account of overinformativity, one which explains differences between languages with prenominal and postnominal adjectives (Paula Rubio-Fernandez, 2020).

But overinformativity has its limits: There is

Proceedings of the Probability and Meaning Conference (PaM 2020), pages 104–108
October 14-15 2020 ©2020 Association for Computational Linguistics

still a basic trade-off between accuracy and cost at work in the realm of referring expressions. This basic premise predicts that referring expressions for objects in scenes with multiple objects of the same type will tend to be longer, as more content is necessary in order to distinguish one referent from another.

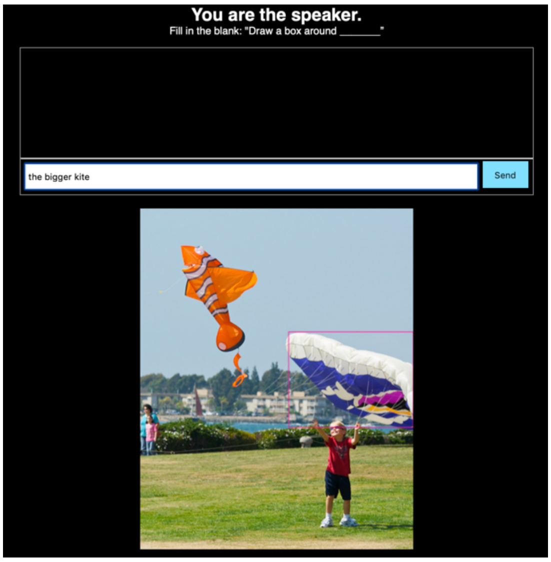
Captions are not subject to the same pressures. The purpose of a caption is not to distinguish one object from another, but rather to describe what is in the picture. Hence we predict that the number of objects in a scene with the same type should have a significant impact on the length of a referring expression for an object of that type, but either less or no impact on the length of a caption.

As we will show, this prediction is borne out by the data. We find furthermore that captions generally involve indefinite descriptions while referring expressions use definite descriptions, and referring expressions typically make use of more relational vocabulary (e.g. *left*, *closest*) than captions.

3 Approach

The Visual Genome corpus (Krishna et al., 2017) provides a set of captions for objects in complex scenes, called *region descriptions*. We selected a subset of these images in order to construct a dataset of corresponding referring expressions. Our dataset was constructed based on object types (e.g. horse, phone, vase) such that there exist images with one, two, and three objects of that type (e.g. one horse, two horses, and three horses). For each of the types satisfying this condition, we included two images with a SINGLE instance of the type, two with two instances (DOUBLE), and two with three instances of the type (TRIPLE). A total of 198 images were included, comprising 33 sextuples.

We developed an interactive web-based reference game in which a speaker was matched with a listener, and was told to complete the sentence *Draw a box around* _____, for an object in a complex scene designated with a bounding box (see Figure 1). Participants were randomly assigned the role of speaker or listener and communicated through a modified chat window. The listener was instructed to draw a box around the entity indicated by the speaker, and the box drawn by the listener was shown to the speaker as feedback. We filtered out participants who did not attempt to distinguish one object from another in their responses (e.g. referring to one of three teddy bears as 'toy'), and we normalized the responses, taking into account self-corrections and variations in how speakers interpreted the task (e.g. 'the hose |I mean the horse' was normalized to 'the horse', and 'Draw a box around the center horse' was normalized to 'the center horse').



You are the speaker.

Fill in the blank: "Draw a box around _______"

the bigger kite

Figure 1: Speaker's point of view in reference game.

Our predictions about length are conditional on whether the referring expressions use a synonym or the same word as the target type; a hyponym would be an alternative strategy to include more specific information. We therefore analyzed the sense relation between the head noun of the description and the target type noun. We used a dependency parser to identify the head noun of the description, and categorized the head noun as a HYPONYM, SYNONYM (or SAME word), or HYPERNYM of the noun corresponding to the target type using WordNet[1].

We then carried out an analysis of the external syntax of these *semi-normalized responses*. Some participants used full definite descriptions, as in *the horse in the middle*, while others left off the initial definite article *horse in the middle*, and others used an even more telegraphic style: *horse in middle*. The variation in style is of interest in its own right, but also makes the descriptions difficult to compare in terms of length. To resolve this, we normalized the responses to make them full noun phrases. We compared the length of the resulting *fully normalized responses*, comparing them to the captions in Visual Genome for the corresponding

[1] https://wordnet.princeton.edu

regions.

4 Results and Discussion

Sample results are shown in Figure 3. In the image with a **single (salient) plane**, over-informative adjectives (*red and white*) are provided to describe the unique salient plane in the image (there is in fact another one in the background), while the referring expression provides just enough information to identify the salient plane (*the plane*). In the image with **three polar bears**, the caption is shorter than the referring expression; the caption simply describes the entity as a polar bear, while the referring expression provides enough information to distinguish the entity from other ones in the scene (the negation of a relational property, *getting licked*). In the image with **two horses**, the caption and the referring expression are of comparable length, but the caption provides non-distinguishing information; the referring expression uses the relational expression *darker* to uniquely identify a referent. In the image with **three planes**, again the caption and the referring expression are of comparable length, and the caption contains enough information to distinguish the referent from the other potential referents in the scene. However, the referring expression uses the relational term *middle*, while the caption describes a non-relational attribute of the object. And of course, the referring expressions use definite articles, while the captions tend to use indefinite articles. These images are representative of the overall set of patterns.

Let us turn now to a quantitative analysis. We note first that the overwhelming majority of the referring expressions we gathered (**94.5%**) were noun phrases headed by the same noun as the target type or a synonym; only 5.5% were a hyponym or a hypernym. We therefore predict for our dataset overall that in images with multiple instances of a given type, referring expressions picking out one of those instances should be longer, in comparison to images with only a single instance of the type.

Of the unique **referring expressions** we gathered, **63% used a definite description**. **Less than 1% used an indefinite description.** The remaining group was predominantly made up of descriptions lacking an initial article, e.g. *horse on (the) left*, with only a handful of exceptions. In contrast, in the corresponding **region descriptions (captions)**, **4.7% used a definite description**, and **39.6% used an indefinite description**. The remaining set were predominantly noun phrases with no initial article (e.g. *large brown bear by a rocky wall*; notice here that the embedded noun phrase is indefinite, however). Perhaps surprisingly, 11.9% of the region descriptions took the form of a sentence, e.g. *Pizza is thin crust* or *The zebra has stripes*. It seems the region description data reflects a range of approaches to the annotation task; this is a source of noise in the data.

We now compare the captions to the referring expressions with respect to length. The results are summarized in Table 1, which shows the mean length of utterances for both region descriptions (captions) and referring expressions, by number of objects of the same type within the image. These results are also visualized in Figure 2, which shows the distribution of lengths (note that the points are jittered, so as to avoid overplotting).

These results support the hypothesis that referring expressions and captions are subject to very different pressures with respect to informativity. Referring expressions include descriptive information for the purpose of distinguishing one referent from another, while captions do not.

Finally, the kind of information that can help to discriminate among referents often consists in relations that instances of the type stand in to each other (e.g. *darker brown*, *in the middle*, *closest*, *on the right*). We defined a *relational modifier* narrowly as a modifier that specifies a characteristic of an object in relation to another instance of the type named by the head noun, excluding gradable size adjectives like *big*. Even on this narrow definition, we find a strong difference between captions and referring expressions, with **captions** exhibiting such modifiers at a rate of **less than one percent**, and **referring expressions** exhibit-

	REF. EXP	CAPTION
SINGLE	2.95 ($sd = 2.2$)	4.45 ($sd = 1.8$)
DOUBLE	4.84 ($sd = 2.9$)	4.46 ($sd = 2.9$)
TRIPLE	5.35 ($sd = 2.7$)	3.45 ($sd = 2.9$)
t	2.1	-0.66
$P(> \|t\|)$	0.039 (*)	0.511 (n.s.)

Table 1: Mean length in words for captions vs. referring expressions, along with t statistics and P-values for OLS-based linear regression models estimating the effect of target type count on length.

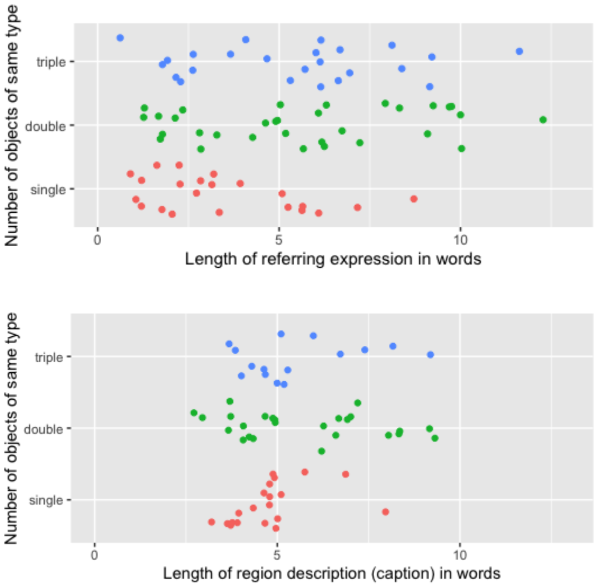


Figure 2: Effect of number of instances of target type on length for referring expressions (top) and captions (bottom) (points jittered).

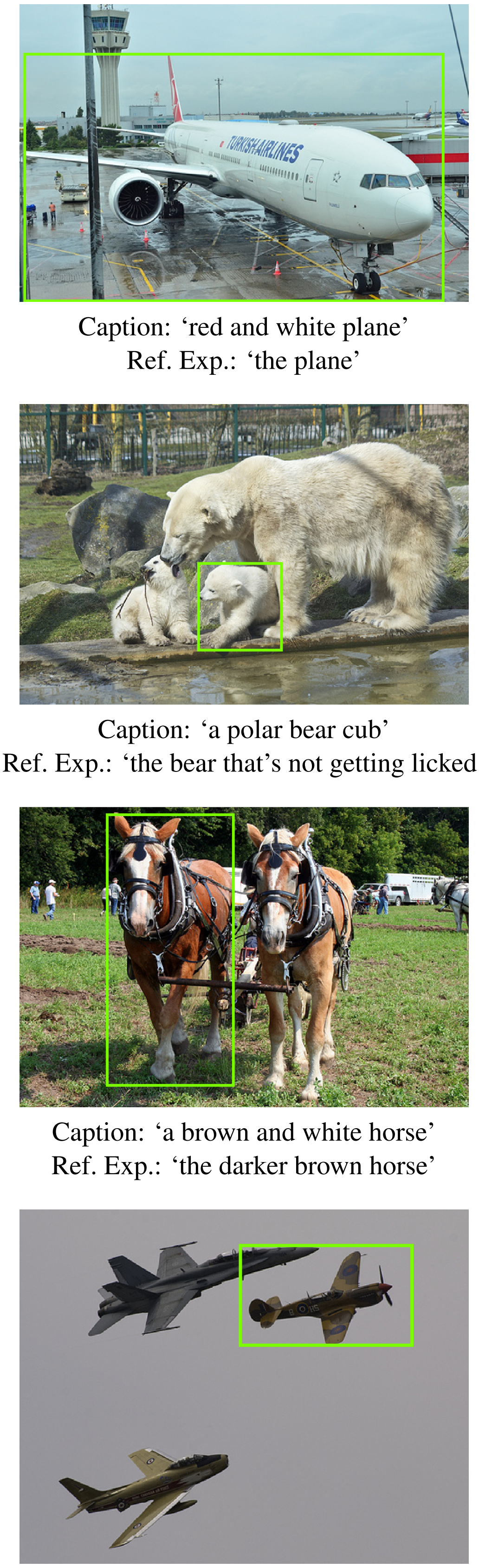
ing them at a rate of **26.3%**.

5 Conclusion

This comparison has shown that referring expressions and captions are subject to very different pressures with respect to informativity. When there is only a single instance of a given type (or only one instance that is visually salient), then it suffices to refer to it using 'the [noun]', where '[noun]' identifies the type. A caption, on the other hand, is there to tell someone about the object, so descriptive detail is more likely to be added even when it does not help to identify the referent.

But captions are not systematically longer than referring expressions, either. Descriptive modifiers will be added to a referring expression when they serve the purpose of distinguishing the referent from other ones, i.e., when they are informative. This is why expressions referring to objects of a type that is multiply instantiated within a scene tend to be longer. A caption and the corresponding referring expression may also be equally long, but the kind of information they contain is different: a caption is more likely to contain information that does not help to discriminate among the possible referents. Relational vocabulary is for distinguishing among referents.

We hope that these findings will enable image captioning datasets to be leveraged more effectively in systems for generating expressions that



Caption: 'red and white plane'
Ref. Exp.: 'the plane'

Caption: 'a polar bear cub'
Ref. Exp.: 'the bear that's not getting licked'

Caption: 'a brown and white horse'
Ref. Exp.: 'the darker brown horse'

Caption: 'plane with a propeller on the front'
Ref. Exp.: 'the airplane in the middle'

Figure 3: Captions versus referring expressions for selected images.

refer to objects in complex scenes.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Peter Baumann, Brady Clark, and Stefan Kaufmann. 2014. Overspecification and the cost of pragmatic reasoning about referring expressions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1898–1903.

Reuben Cohn-Gordon, Noah Goodman, and Chris Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, NAACL-HLT*, volume 2, pages 439–443.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. 2019. An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics*, volume 2, pages 81–90.

Judith Degen, Robert X. D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2019. When redundancy is rational: A bayesian approach to 'overinformative' referring expressions. *CoRR*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Albert Gatt, Roger P. G. van Gompel, Kees van Deemter, and Emiel Krahmer. 2013. Are we bayesian referring expression generators? In *Proceedings of the Cogsci workshop on Production of Referring expressions. Associated with the 35th Annual Conference of the Cognitive Science Society*, Berlin.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehensions of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Julian Jara-Ettinger Paula Rubio-Fernandez, Francis Mollica. 2020. Why searching for a blue triangle is different in english than in spanish.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.