

Know What You Don’t Know: Modeling a Pragmatic Speaker that Refers to Objects of Unknown Categories

Sina Zarrieß

Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
sina.zarriess@uni-bielefeld.de

David Schlangen

Linguistics Department
University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

Zero-shot learning in Language & Vision is the task of correctly labelling (or *naming*) objects of novel categories. Another strand of work in L&V aims at pragmatically informative rather than “correct” object descriptions, e.g. in reference games. We combine these lines of research and model *zero-shot reference games*, where a speaker needs to successfully refer to a novel object in an image. Inspired by models of “rational speech acts”, we extend a neural generator to become a pragmatic speaker reasoning about uncertain object categories. As a result of this reasoning, the generator produces fewer nouns and names of distractor categories as compared to a literal speaker. We show that this conversational strategy for dealing with novel objects often improves communicative success, in terms of resolution accuracy of an automatic listener.

1 Introduction

It is commonly agreed that even massive resources for language & vision (Deng et al., 2009; Chen et al., 2015; Krishna et al., 2017) will never fully cover the huge range of objects to be found “in the wild”. This motivates research in zero-shot learning (Lampert et al., 2009; Socher et al., 2013; Hendricks et al., 2016), which aims at predicting correct labels or names for objects of novel categories, typically via external lexical knowledge such as, e.g., word embeddings.

More generally, however, uncertain knowledge of the world that surrounds us, including novel objects, is not only a machine learning challenge: it is simply a very common aspect of human communication, as speakers rarely have perfect representations of their environment. Precisely the richness of verbal interaction allows us to communicate these uncertainties and to collaborate towards communicative success (Clark and Wilkes-Gibbs, 1986). Figure 1 illustrates this general



refexp: right thingy



refexp: left blue

Figure 1: RefCOCO expressions referring to difficult/unknown objects

point with two examples from the RefCOCO corpus (Yu et al., 2016), providing descriptions of visual objects from an interactive reference game. Here, the use of the unspecific *thingy* and the omission of a noun in *left blue* can be seen as pragmatically plausible strategies that avoid confusing the listener with potentially inaccurate names for difficult-to-name objects. While there has been a lot of recent and traditional research on pragmatically informative object descriptions in reference games (Mao et al., 2016; Yu et al., 2017; Cohn-Gordon et al., 2018; Dale and Reiter, 1995; Frank and Goodman, 2012), conversational strategies for dealing with uncertainties like novel categories are largely understudied in computational pragmatics, though see, e.g., work by Fang et al. (2014).

In this paper, we frame zero-shot learning as a challenge for pragmatic modeling and explore *zero-shot reference games*, where a speaker needs to describe a novel-category object in an image to an addressee who may or may not know the category. In contrast to standard reference games, this game explicitly targets a situation where relatively common words like object names are likely to be more inaccurate than other words like e.g. attributes. We hypothesize that Bayesian reasoning in the style of Rational Speech Acts, RSA (Frank and Goodman, 2012), can extend a neural generation model trained to refer to objects of known categories, towards zero-shot learning. We im-

plement a Bayesian decoder reasoning about categorical uncertainty and show that, solely as a result of pragmatic decoding, our model produces fewer misleading object names when being uncertain about the category (just as the speakers did in Figure 1). Furthermore, we show that this strategy often improves reference resolution accuracies of an automatic listener.

2 Background

We investigate referring expression generation (REG henceforth), where the goal is to compute an utterance u that identifies a target referent r among other referents R in a visual scene. Research on REG has a long tradition in natural language generation (Krahmer and Van Deemter, 2012), and has recently been re-discovered in the area of Language & Vision (Mao et al., 2016; Yu et al., 2016; Zarrieß and Schlangen, 2018). These latter models for REG essentially implement variants of a standard neural image captioning architecture (Vinyals et al., 2015), combining a CNN and an LSTM to generate an utterance directly from objects marked via bounding boxes in real-world images.

Our approach combines such a neural REG model with a reasoning component that is inspired by theory-driven Bayesian pragmatics and RSA (Frank and Goodman, 2012). We will briefly sketch this approach here. The starting point in RSA is a model of a “literal speaker”, $S_0(u|r)$, which generates utterances u for the target r . The “pragmatic listener” L_0 then assigns probabilities to all referents R based on the model S_0 :

$$L_0(r|u) \propto \frac{S_0(u|r) * P(r)}{\sum_{r_i \in R} S_0(u|r_i) * P(r_i)} \quad (1)$$

In turn, the “pragmatic speaker” S_1 reasons about which utterance is more discriminative and will be resolved to the target by the pragmatic listener:

$$S_1(u|r) \propto \frac{L_0(r|u) * P(u)}{\sum_{u_i \in U} L_0(r|u_i) * P(u_i)} \quad (2)$$

(S_0 and L_0 are components of the recursive reasoning of S_1 and not in fact separate agents.)

There has been some previous work on leveraging RSA-like reasoning for neural language generation. For instance, Cohn-Gordon et al. (2018) implement the literal speaker as a neural captioning model trained on non-discriminative image descriptions. On top of this neural semantics, they

build a pragmatic speaker that produces more discriminative captions, applying equation 2 at each step of the inference process. They evaluate their model in a reference game where an automatic listener (trained on a different portion of the image data) is used to test whether the generated caption singles out the target image among a range of distractor images. A range of related articles have extended neural captioning models with decoding procedures geared towards vocabulary expansion (Anderson et al., 2017; Agrawal et al., 2018) or contextually discriminative scene descriptions (Andreas and Klein, 2016; Vedantam et al., 2017).

Previous work on REG commonly looks at visual scenes with multiple referents of identical or similar categories. Here, speakers typically produce expressions composed of a head noun, which names the category of the target, and a set of attributes, which distinguish the target from distractor referents of the same category (Krahmer and Van Deemter, 2012). Our work adds an additional dimension of uncertainty to this picture, namely a setting where the category of the target itself might not be known to the model and, hence, cannot be named with reasonable accuracy. In this setting, we expect that a literal speaker (e.g. a neural REG model trained for a restricted set object categories) generates misleading references, e.g. containing incorrect head nouns, as it has no means of “knowing” which words risk being inaccurate for referring to novel objects. The following Section 3 describes how we modify the RSA approach for reasoning in such a zero-shot reference game.

3 Model

Inspired by the approach in Section 2, we model our pragmatic zero-shot speaker as a neural generator (the literal speaker) that is decoded via a pragmatic listener. In contrast to the listener in Equation (1), however, our listener possesses an additional latent variable C , which reflects its beliefs about the target’s category. This hidden belief distribution will, in turn, allow the pragmatic speaker to reason about how accurate the words produced by the literal speaker might be.

Our Bayesian listener will assign a probability $P(r|u)$ to a referent r conditioned on the utterance u by the (literal) speaker. To do that, it needs to calculate $P(u|r)$, as in Equation 1. While previous work on RSA typically equates $P(u|r)$ with $S_0(u|r)$, we are going to modify the way this prob-

ability is calculated. Thus, we assume that our listener has hidden beliefs about the category of the referent, that we can marginalize over as follows:

$$\begin{aligned}
P(u|r) &= \sum_{c_i \in C} P(u, c_i|r) = \sum_{c_i \in C} \frac{P(u, c_i, r)}{P(r)} \\
&= \sum_{c_i \in C} \frac{P(r) * P(c_i|r) * P(u|c_i, r)}{P(r)} \\
&\propto \sum_{c_i \in C} P(c_i|r) * P(u|c_i)
\end{aligned} \tag{3}$$

As a simplification, we condition u only on c_i , instead of $P(u|c_i, r)$. This will allow us to estimate $P(u|c_i)$ directly via maximum likelihood on the training data, i.e. in terms of word probabilities conditioned on categories (observed in training). The pragmatic listener is defined as follows:

$$\begin{aligned}
L_0(r|u) &= \frac{P(u|r) * P(r)}{P(u)} \\
&\propto \sum_{c_i \in C} P(c_i|r) * P(u|c_i)
\end{aligned} \tag{4}$$

For instance, let's consider a game with 3 categories and two words, the less specific *left* with $P(u|c_i) = \frac{1}{2}$ for all $c_i \in C$ and the more specific *bus* with $P(u|c_1) = \frac{9}{10}, P(u|c_2) = \frac{1}{10}, P(u|c_3) = \frac{1}{10}$. When the listener is uncertain and predicts $P(c_i|r) = \frac{1}{3}$ for all $c_i \in C$, this yields $L_0(r|\text{left}) = 0.5$ and $L_0(r|\text{bus}) = 0.36$, meaning that the less specific *left* will be more likely resolved to the target r . Vice versa, when the listener is more certain, e.g. $P(c_1|r) = \frac{9}{10}, P(c_2|r) = \frac{1}{10}, P(c_3|r) = \frac{1}{10}$, more specific words will be preferred: $L_0(r|\text{bus}) = 0.83$ and $L_0(r|\text{left}) = 0.55$.

The definition of the pragmatic speaker is straightforward:

$$S_1(u|r) = S_0(u|r) * L_0(r|u)^\alpha \tag{5}$$

Intuitively, S_1 guides its potentially over-optimistic language model (S_0) to be more cautious in producing category-specific words, e.g. nouns. The idea is that the degree to which a word is category-specific and, hence, risky in a zero-shot reference game can be determined on descriptions for objects of *known* categories and is expressed in $P(u|c)$. For *unknown* categories, the pragmatic speaker can deliberately avoid these

category-specific words and resort to describing other visual properties like colour or location.¹

Similar to Cohn-Gordon et al. (2018), we use incremental, word-level inference to decode the pragmatic speaker model in a greedy fashion:

$$S_1^t(w|r, u_{t-1}) = S_0^t(w|r, u_{t-1}) * L_0(r|w)^{\alpha+\beta} \tag{6}$$

At each time step, we generate the most likely word determined via S_0 and L_0 . The parameters α and β will determine the balance between the literal speaker and the listener. While α is simply a constant (set to 2, in our case), β is zero as long as w does not occur in u_{t-1} and increases when it does occur in u_{t-1} (it is then set to 2). This ensures that there is a dynamic tradeoff between the speaker and the listener, i.e. for words that occur in previously generated utterance prefix, the language model probabilities (S_0) will have comparatively more weight than for new words.

4 Exp. 1: Referring without naming?

Section 3 has introduced a model for referring expression generation (REG) in a zero-shot reference game. This model, and its pragmatic decoding component in particular, is designed to avoid words that are specific to categories when there is uncertainty about the category of a target object, in favour of words that are not specific to categories like, e.g., colour or location attributes. In the following evaluation, we will test how this reasoning component actually affects the referring behavior of the pragmatic speaker as compared to the literal speaker, which we implement as neural supervised REG model along the lines of previous work (Mao et al., 2016; Yu et al., 2016). As object names typically express category-specific information in referring expressions, we focus the comparison on the nouns generated in the systems' output.

4.1 Training

Data We conduct experiments on RefCOCO (Yu et al., 2016) referring expressions to objects in MSCOCO (Lin et al., 2014) images. As is commonly done in zero-shot learning, we manually select a range of different categories as targets for our zero-shot game, cf. (Hendricks et al., 2016). Out of the 90 categories

¹We leave it for future work to combine this approach with a listener reasoning about distractor objects in the scene (as in Equation 1).

in MSCOCO, we select 6 medium-frequent categories (*cat, horse, cup, bottle, bus, train*), that are similar to those in (Hendricks et al., 2016). For each category, we divide the training set of RefCOCO into a new train-test split such that all images with an instance of the target zero-shot category are moved to the test set.

Generation Model (S_0) We implement a standard CNN-LSTM model for REG, trained on pairs of image regions and referring expressions. The architecture follows the baseline version of (Yu et al., 2016). We crop images to the target region, and obtain the fc features from VGG (Simonyan and Zisserman, 2014). We set the word embedding layer size to 512, and the hidden state to 1024. We optimized with ADAM, set the batch size to 32 and the learning rate to 0.0004. The number of training epochs is 5 (verified on the RefCOCO validation set).

Uncertainty Estimation Similar to previous work in zero-shot learning, we factor out the problem of automatically determining the model’s certainty with respect to an object’s category, cf. (Lampert et al., 2009; Socher et al., 2013): for computing $L_0(r|u)$, we set $P(c_i|r)$ to be a uniform distribution over categories, meaning that the model is maximally uncertain about the referent’s category. We leave exploration of a more realistic uncertainty or novelty prediction to future work.

4.2 Evaluation

Measures We test to what extent our models produces incorrect names for novel objects. First, for each zero-shot category, we define a set of distractor nouns (**distr-noun**), which correspond to the names of the remaining categories in MSCOCO. Any choice of noun from that set would be wrong, as the categories are pairwise disjunct; the exploration of other nouns (e.g. *thingy, animal*) is left for future work. In Table 1, “% distr-noun” refers to how many expressions generated for an instance of a zero-shot category contain such an incorrect distractor noun. Second, we count how many generated expressions do not contain any noun (**no-noun**) at all, according to the NLTK POS tagger.

Results Table 1 shows that the proportion of output expressions containing a distractor noun *decreases* markedly from S_0 to S_1 , whereas the proportion of expression without any name *increases*

Model		% distr-noun	% no-noun
<i>cat</i>	S_0	0.606	0.107
	S_1	0.484	0.193
<i>horse</i>	S_0	0.683	0.085
	S_1	0.572	0.30
<i>cup</i>	S_0	0.627	0.079
	S_1	0.332	0.172
<i>bottle</i>	S_0	0.398	0.275
	S_1	0.166	0.562
<i>bus</i>	S_0	0.743	0.066
	S_1	0.612	0.247
<i>train</i>	S_0	0.759	0.166
	S_1	0.558	0.37

Table 1: Names and nouns contained in generation output for two speakers (S_0, S_1)



Target (unknown cat):
left horse
 S_0 : left person ✗
 S_1 : left black ✓

Figure 2: Qualitative Example

markedly from S_0 to S_1 . First of all, this suggests that our baseline model S_0 does, in many cases, not know what it does not know, i.e. it is not aware that it encounters a novel category and frequently generates names of known categories encountered during training. However, even in this simple model, we find a certain portion of output expressions that do not contain any name (e.g. 27% for *bottle*, but only 6% for *bus*). The results also confirm our hypothesis that the pragmatic speaker S_1 avoids to produce “risky” or specific words that are likely to be confused for uncertain or unknown categories. It is worth stressing here that this behaviour results entirely from the Bayesian reasoning that S_1 uses in decoding; the model does not have explicit knowledge of linguistic categories like nouns, names or other taxonomic knowledge.

5 Exp. 2: Communicative success

The Experiment in Section 4 found that the pragmatic speaker uses less category-specific vocabulary when referring to objects of novel categories as compared to a literal speaker. Now, we need to establish whether the resulting utterances still achieve communicative success in the zero-shot reference game, despite using less specific vocab-

Zero-shot category	Similar category
<i>cat</i>	dog, cow
<i>horse</i>	dog, cow
<i>cup</i>	bowl, bottle, wine glass
<i>bottle</i>	vase, wine glass
<i>bus</i>	car, train, truck
<i>train</i>	car, bus, truck

Table 2: Target and distractor categories used for testing in Exp. 2

ulary (as shown above). We test this automatically using a model of a “competent” listener, that knows the respective object categories. This is supposed to approximate a conversation between a system and a human that has more elaborate knowledge of the world than the system.

The evaluation listener One pitfall of using a trained listener model (instead of a human) for task-oriented evaluation is that this model might simply make the same mistakes as the speaker model as it is trained on similar data. To avoid this circularity, Cohn-Gordon et al. (2018) train their listener on a different subset of the image data. Rather than training on *different* data, we opt for training the listener on *better* data, as we want it to be as strict and human-like as possible. For instance, we do not want our listener model to resolve an expression like *the brown cat* to a *dog*. We train S_{eval} as a neural speaker on the entire training set and give L_{eval} access to ground-truth object categories. The ground-truth category c_r of a referent r is used to calculate $P(n_u|c_r)$ where n_u is the object name contained in the utterance u . $P(n_u|c_r)$ is estimated on the entire training set.

$$L_{eval}(r|u, c_r) = S_{eval}(u|r) * P(n_u|c_r) \quad (7)$$

$P(n_u|c_r)$ will be close to zero if the utterance contains a rare or wrong name for the category c_r , and L_{eval} will then assign a very low probability to this referent. We apply this listener to all referents in the scene and take the argmax.

Test set The set **TS-image** pairs each target with other (annotated!) objects in the same image, a typical set-up for reference resolution. As many images in RefCOCO only have distractors of the *same* category as the target (which is not ideal for our purposes), we randomly sample an additional test set called **TS-distractors**, pairing zero-

	Model	TS-image	TS-distractors
<i>cat</i>	S_0	0.516	0.343
	S_1	0.603	0.386
<i>horse</i>	S_0	0.644	0.096
	S_1	0.589	0.150
<i>cup</i>	S_0	0.721	0.483
	S_1	0.674	0.540
<i>bottle</i>	S_0	0.502	0.275
	S_1	0.517	0.306
<i>bus</i>	S_0	0.789	0.405
	S_1	0.759	0.361
<i>train</i>	S_0	0.658	0.202
	S_1	0.667	0.305

Table 3: Reference resolution accuracies obtained from listener L_{eval} on expressions by S_0, S_1

shot targets with 4 distractors of a similar category, which we defined manually, shown in Table 2. This is slightly artificial as objects are taken out of the coherent spatial context, but it helps us determining whether our model can successfully refer in a context with similar, but not identical, categories.

Results As shown in Table 3, the S_1 model improves the resolution accuracy for all categories on TS-distractors, except for *bus*. On TS-image, resolution accuracies are generally much higher and the comparison between S_0 and S_1 gives mixed results. We take this as positive evidence that S_1 improves communicative success in a relevant number of cases, but it also indicates that combining this model with the more standard RSA approach could be promising. Figure 2 shows a qualitative example for S_1 being more successful than S_0 .

6 Conclusion

We have presented a pragmatic approach to modeling zero-shot reference games, showing that Bayesian reasoning inspired by RSA can help decoding a neural generator that refers to novel objects. The decoder is based on a pragmatic listener that has hidden beliefs about a referent’s category, which leads the pragmatic speaker to use fewer nouns when being uncertain about this category. While some aspects of the experimental setting are, admittedly, simplified (e.g. compilation of an artificial test set, uncertainty estimation), we believe that this is an encouraging result for scaling models in computational pragmatics to real-world conversation and its complexities.

References

Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2018. nocaps: novel object captioning at scale. *arXiv preprint arXiv:1812.08658*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Reuben Cohn-Gordon, Noah Goodman, and Christopher 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, Volume 2 (Short Papers)*, pages 439–443. Association for Computational Linguistics.

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

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Rui Fang, Malcolm Doering, and Joyce Y Chai. 2014. Collaborative models for referring expression generation in situated dialogue. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

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

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10.

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, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Vision and Pattern Recognition*, pages 951–958. IEEE.

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*, volume 8693, pages 740–755. Springer International Publishing.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR 2016*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Computer Vision and Pattern Recognition (CVPR)*, volume 3.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. *Proceedings of INLG 2018*.