Decoupling Pragmatics: Discriminative Decoding for Referring Expression Generation

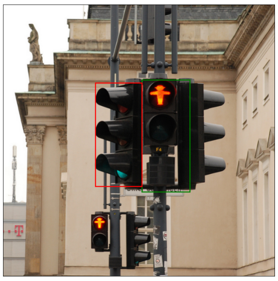
Simeon Schüz and **Sina Zarrieß**
Bielefeld University
{simeon.schuez,sina.zarriess}@uni-bielefeld.de

Abstract

The shift to neural models in Referring Expression Generation (REG) has enabled more natural set-ups, but at the cost of interpretability. We argue that integrating pragmatic reasoning into the inference of context-agnostic generation models could reconcile traits of traditional and neural REG, as this offers a separation between context-independent, literal information and pragmatic adaptation to context. With this in mind, we apply existing decoding strategies from discriminative image captioning to REG and evaluate them in terms of pragmatic informativity, likelihood to ground-truth annotations and linguistic diversity. Our results show general effectiveness, but a relatively small gain in informativity, raising important questions for REG in general.

1 Introduction

In recent years, neural models have become the workhorses for Referring Expression Generation (REG, e.g. Mao et al., 2016; Yu et al., 2016; Zarrieß and Schlangen, 2018), as in other tasks in the Vision and Language (V&L) domain (Mogadala et al., 2019). In REG, this was accompanied by a major shift in how the task was conceptualized. Classical approaches (e.g. Dale, 1989; Dale and Reiter, 1995) mostly investigated rule-based procedures to determine combinations of properties that distinguish target objects from distractors, based on knowledge bases of objects and associated attributes (Krahmer and van Deemter, 2019). Recent work in REG has shifted to more natural settings (e.g. objects in photographs, cf. Figure 1), but at the expense of interpretability: Since continuous representations have replaced knowledge bases as the input, pragmatic processes in neural REG no longer operate on symbolic properties, but are deeply woven into model architectures and training schemes.



Greedy	traffic light
Beam	traffic light
$ES_{\lambda 0.5}$	red light
$RSA_{\alpha 1.0}$	stop light

Figure 1: Example from RefCOCO. ES and RSA describe the target (marked green) less ambiguously.

Decoding and reasoning methods for *discriminative image captioning* (Vedantam et al., 2017; Cohn-Gordon et al., 2018) could represent a middle ground in this regard: During inference, predictions from a captioning model are reranked according to pragmatic principles, boosting contextually informative and inhibiting ambiguous utterances. This offers interesting similarities to traditional REG, as it is carried out through explicit algorithms and targets symbolic representations (e.g. word tokens). *Discriminative decoding* has been shown to be effective for image captioning (Vedantam et al., 2017; Cohn-Gordon et al., 2018; Schüz et al., 2021). In this work, we investigate discriminative decoding for REG, adapting the methods from Vedantam et al. (2017) and Cohn-Gordon et al. (2018). We compare them to standard *greedy* and *beam search* decoding, in terms of informativity, likelihood to ground-truth annotations, and linguistic diversity. We show that discriminative decoding increases informativity and diversity, although the results are less clear than expected. We attribute this, in part, to the way human annotations are collected, highlighting implications for REG research in general.

2 Background

Traditional and neural REG In REG, the goal is to generate descriptions for entities, which al-

Proceedings of the Conference on Reasoning and Interaction (ReInAct), pages 47–52
October 4-6, 2021. ©2021 Association for Computational Linguistics

low their identification in a given context (Reiter and Dale, 2000); i.e. generating expressions with sufficient, but not too much information, following a Gricean notion of pragmatics (Grice, 1975; Krahmer and van Deemter, 2019). In classic work, target and distractor objects were defined in terms of symbolic attributes and associated values (e.g. color - red). The full REG task was conceived as involving different levels of processing, i.e. lexicalization, content selection and surface realization (Reiter and Dale, 2000; Krahmer and van Deemter, 2019). However, foundational work in REG has mostly focused on algorithms for finding *distinguishing* sets of attribute-value pairs, which apply to the target, but *rule out* distractor objects, such as the *Incremental Algorithm* (*IA*, Dale and Reiter, 1995). This algorithm iterates over the attribute set in a pre-defined order, selects an attribute if it rules out objects from the set of distractors and terminates when the set is empty. It has been refined, extended and tested in subsequent work (Krahmer et al., 2003; Mitchell et al., 2010; van Deemter et al., 2012; Clarke et al., 2013).

In recent years, neural models have enabled REG set-ups based on real-world images (Kazemzadeh et al., 2014; Gkatzia et al., 2015; Mao et al., 2015; Zarrieß and Schlangen, 2016, 2018; Tanaka et al., 2019; Liu et al., 2020; Kim et al., 2020; Panagiaris et al., 2020, 2021), representing scenes with many different types of real-world objects. Most commonly, neural REG models follow the encoder-decoder scheme and are trained end-to-end. Based on low-level visual representations as the input, various aspects of the task are modeled jointly, e.g. lexicalization and content selection.

Neural Generation and Pragmatics Various approaches were proposed to generate more discriminative expressions in neural REG, such as specialized training objectives (Mao et al., 2016), enhanced input representations and joint generation for objects in the same scene (Yu et al., 2016), listener / comprehension components or reinforcement modules (Luo and Shakhnarovich, 2017; Yu et al., 2017), and classifiers which predict attributes for depicted objects (Liu et al., 2017, 2020). Here, the REG models are trained to jointly determine truthful descriptions for depicted objects and formulate expressions that are unambiguous in a given context. Hence, semantic and pragmatic processing are tightly intertwined, preventing a clear separation between context-independent information and pragmatic adaption as in traditional REG.

In image captioning, e.g. Andreas and Klein (2016); Vedantam et al. (2017); Cohn-Gordon et al. (2018) tried to generate pragmatically informative captions, by decoding general captioning models, at testing time, to produce captions that discriminate target images from a given set of distractor images. This corresponds more closely to approaches such as the IA, as it takes place over a finite set of symbolic (word) tokens and leaves the *literal* generation process untouched. In this work we use the methods proposed by Vedantam et al. (2017) and Cohn-Gordon et al. (2018) and adapt them to neural REG. For evaluation, we roughly follow the experimental set-up from Schüz et al. (2021) and consider likelihood to human annotations, informativity and diversity, the latter as a proxy for the degree of linguistic adaptation to context.

3 Experiments

3.1 Decoding Methods

We compare contrasting decoding methods, which focus either on likelihood or informativity.

For the former, **Greedy Search** selects the token with the highest probability at each time step. **Beam Search** simultaneously extends a fixed number of k hypotheses at each step (here: $k = 5$).

Based on the *Rational Speech Acts* model (Frank and Goodman, 2012), **RSA Decoding** (Cohn-Gordon et al., 2018, henceforth *RSA*) aims for higher informativity by integrating pragmatic reasoning into the iterative unrolling of recurrent captioning models. Given a target and a set of distractors, the *literal speaker* S_0 generates initial distributions over possible next tokens. The *literal listener* L_0 determines which tokens effectively distinguish the target from the distractors. Finally, the *pragmatic speaker* S_1 selects tokens rated informative by L_0. A rationality parameter α specifies the relative influence of L_0 in S_1, cf. Cohn-Gordon et al. (2018) for more details. In our REG setting, targets and distractors are objects in the same image.

In the conceptually similar **Emitter-Suppressor** approach (Vedantam et al., 2017, henceforth *ES*), a speaker (*emitter*) models a caption for a target image I_t in conjuction with a listener function (*suppressor*) that rates the discriminativeness of the utterance with regard to a distractor image, cf. Vedantam et al. (2017). λ is a rationality parameter; the smaller the value of λ, the more the suppressor is weighted. We adapt the extended implementation

	testA			testB			testA+			testB+			testg		
	BL$_1$	CDr	det.	BL$_1$	CDr	det.	BL$_1$	CDr	det.	BL$_1$	CDr	det.	BL$_1$	CDr	det.
greedy	**53.2**	0.79	75.1	**56.2**	1.27	66.4	**45.6**	**0.68**	63.8	**33.9**	0.64	43.0	45.0	0.71	54.4
beam	52.8	**0.81**	75.0	55.4	**1.31**	66.0	40.4	0.66	62.4	32.5	**0.75**	43.6	**45.0**	**0.79**	54.8
ES$_{\lambda 0.7}$	52.9	0.80	77.6	55.7	1.24	70.5	40.3	0.67	67.1	30.9	0.71	46.9	44.1	0.74	57.1
ES$_{\lambda 0.5}$	49.8	0.73	**80.6**	53.3	1.11	**71.8**	37.1	0.60	**68.7**	27.4	0.61	47.3	40.4	0.63	57.0
ES$_{\lambda 0.3}$	35.8	0.53	79.1	44.5	0.81	71.3	23.1	0.37	66.5	21.0	0.40	**48.1**	29.5	0.37	54.6
RSA$_{\alpha 0.5}$	50.2	0.77	76.0	55.1	1.26	69.4	32.7	0.58	62.8	28.2	0.67	44.4	43.5	0.70	55.7
RSA$_{\alpha 1.0}$	50.4	0.77	76.4	54.8	1.22	69.1	33.3	0.59	63.2	27.5	0.65	44.6	43.0	0.69	56.1
RSA$_{\alpha 5.0}$	50.9	0.75	79.3	52.7	1.05	70.9	35.4	0.57	66.3	25.7	0.58	46.8	40.2	0.61	**57.4**
human	-	-	84.4	-	-	74.2	-	-	72.6	-	-	57.9	-	-	63.7

Table 1: Likelihood (BLEU$_1$, CIDEr) and informativity (detection) for decoding strategies and data splits.

from Schüz et al. (2021) for multiple distractors.

We use both RSA and ES in a beam search decoding scheme. Whereas in the original approaches the number of distractors is fixed, it varies between images in our REG setting: For an image with n objects, the number of distractors is $n - 1$.

3.2 Data and Model

We use the data and pre-defined splits from RefCOCO, RefCOCO+ (Kazemzadeh et al., 2014) and RefCOCOg (Mao et al., 2016) for training and evaluation. All of these datasets contain English referring expressions to objects in images from MSCOCO (Lin et al., 2014), collected in interactive (RefCOCO, RefCOCO+) or non-interactive (RefCOCOg) settings. In RefCOCO and RefCOCO+, *testA* contains references to humans and *testB* references to other object types. Since both targets and distractors are required for RSA and ES, we remove images from our test splits which contain only a single object. After this, our test splits comprise approximately 1950 (*testA / testA+*), 1750 (*testB / testB+*) and 4000 (*testg*) objects.

We adopted the image captioning model from Lu et al. (2017)[1] as the basis for our REG model. Similar to e.g. Mao et al. (2016), we complemented the original model by supplying 7 location features along with the input image.

3.3 Evaluation

Likelihood is measured through BLEU$_1$ (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015) scores, calculated using the RefCOCO API[2].

For **Diversity**, we calculate the type-token ratio (TTR) for unigrams and bigrams, and the proportion of the model vocabulary used (coverage). Importantly, we look at global diversity, i.e. the corpus-level variation in the usage of words and phrases (van Miltenburg et al., 2018).

Informativity is assessed through the precision of a separate, pre-trained Referring Expression Comprehension model (Luo et al., 2020). Given a generated expression and corresponding image, the model predicts a bounding box which locates the described object in the image. As in the original paper, predictions are deemed correct if the *intersection over union* between predicted and ground-truth bounding boxes is greater than 0.5. Previous work in neural REG mostly assessed informativity through human evaluation (e.g. Yu et al., 2016, 2017; Liu et al., 2017, 2020). We decided for automatic evaluation for the sake of better comparability and exhaustive coverage of our expressions.

4 Results

4.1 Likelihood and Diversity

The results in Table 1 show that discriminative decoding leads to a decrease in both BLEU (BL_1) and CIDEr (CDr). Whereas ES with $\lambda = 0.7$ achieves comparable results to greedy and beam search, both metrics drop if rationality is increased. In most cases, this also applies to RSA. This corresponds to the general findings in Schüz et al. (2021): With higher rationality, ES and RSA generate expressions that deviate further from the model predictions, resulting in lower n-gram overlap.

Similarly, the diversity results in Table 2 confirm the findings in Schüz et al. (2021). Discriminative decoding increases TTR (T_1, T_2) and coverage ($cov.$), indicating that pragmatic reasoning leads to more variation and the usage of a larger vocabulary.

4.2 Informativity

For informativity, ES and RSA outperform greedy and beam search (cf. Table 1, $det.$), although

[1] https://github.com/yufengm/Adaptive

[2] https://github.com/lichengunc/refer

	testA			testB			testA+			testB+			testg		
	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.
greedy	7.0	29.2	4.1	10.6	46.5	5.6	8.7	28.2	4.2	16.5	48.1	7.7	16.7	40.0	10.7
beam	6.6	27.8	3.6	10.3	48.6	5.2	9.4	32.7	4.1	16.4	56.4	6.9	17.5	42.4	10.4
$\text{ES}_{\lambda 0.7}$	8.3	35.0	4.8	11.6	49.7	6.1	12.5	42.6	5.4	19.3	64.0	7.9	19.8	48.0	13.0
$\text{ES}_{\lambda 0.5}$	11.6	46.4	6.9	13.4	54.4	7.6	16.1	54.4	7.9	22.2	70.3	9.5	22.7	56.4	16.5
$\text{ES}_{\lambda 0.3}$	**17.5**	**60.7**	**12.6**	**18.4**	**62.6**	**13.0**	**22.7**	**65.8**	**13.2**	**29.2**	**80.3**	**14.8**	**28.7**	**71.0**	**23.6**
$\text{RSA}_{\alpha 0.5}$	7.8	33.7	4.3	11.6	50.2	5.9	12.1	43.2	5.0	18.5	61.0	7.4	19.8	48.1	12.6
$\text{RSA}_{\alpha 1.0}$	7.8	34.6	4.4	11.8	51.6	5.9	13.0	47.0	5.7	19.7	64.7	7.9	20.4	49.8	13.6
$\text{RSA}_{\alpha 5.0}$	9.4	39.0	5.7	13.5	55.8	7.2	16.1	54.1	7.6	23.9	74.4	10.6	22.5	57.7	16.3
human	24.9	71.7	22.4	27.6	79.6	23.1	31.2	80.6	20.9	39.6	91.2	26.2	34.0	77.8	44.4

Table 2: Diversity results (TTR_1, TTR_2, coverage) for decoding strategies and data splits

even greedy decoding can perform well in certain cases (e.g. testA). Overall, the gain is rather modest: Here, the maximum relative increase is 10% (testB+), whereas Schüz et al. (2021) report more than 30% increase in retrieval[3]. This could be due to upper bounds for possible detection results: Depending on the data set, the comprehension task itself poses a considerable challenge, as can be seen, for example, in the detection results for human annotations in testB+.

An alternative explanation can be seen in the data used to train the model: Unlike in image captioning, the utterances in our datasets were explicitly produced for distinguishing targets and distractors. Thus, by re-using linguistic patterns from the training data, our model might be able to generate relatively informative expressions without even considering the situational context. This way of implicitly learning to fulfill pragmatic requirements might render additional layers of pragmatic reasoning less effective. This hypothesis is supported by the decent results for greedy search e.g. in *testA*. Also, beam search occasionally improves the greedy results, indicating that optimizing model predictions increases pragmatic informativity. In similar set-ups for image captioning, beam search was reported to decrease informativity (Schüz et al., 2021).

Somewhat surprisingly, ES with $\lambda = 0.5$ mostly obtains better results than $\lambda = 0.3$, i.e. higher rationality does not always lead to higher informativity. Figure 2 shows this for a wider range of λ: For every data split, the detection results drop drastically if λ approaches 0. We attribute this to ES struggling to generate well-formed descriptions for high rationalities (as reflected in BLEU_1 and CIDEr for $\lambda = 0.3$), and increasingly diverging from the data used for training the comprehension model.

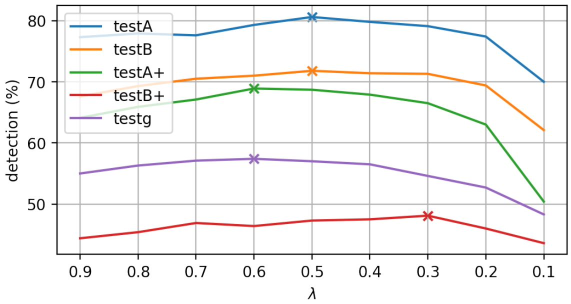


Figure 2: Detection results for different λ settings in ES decoding. Crosses mark the highest detection results.

5 Discussion and Conclusion

Discriminative decoding is appealing for REG, as it combines traits from traditional REG with neural generation models. Our results confirm findings previously reported for image captioning: Discriminative decoding decreases likelihood to ground-truth annotations, but increases informativity and diversity. For informativity, the margin of gain was surprisingly low in our experiments. We attributed this, in part, to the high informativity of underlying model expressions. While this is of importance especially in our setup (both ES and RSA assume a basis of pragmatically neutral descriptions), the question of whether pragmatic informativity has to be explicitly modelled or is implicitly learned from the data is relevant for REG research in general, and should be investigated in subsequent work. To this end, controlling the impact of pragmatic processing as in discriminative decoding could be a valuable instrument.

Beyond this, future work should investigate discriminative decoding in REG in more detail, e.g. to see whether pragmatic reasoning leads to the generation of different or more specific attributes.

[3] Due to differences in tasks, models, data and evaluation, this comparison should be taken with caution.

References

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.

Alasdair DF Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: the influence of visual salience on referring expression generation. *Frontiers in psychology*, 4.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Robert Dale. 1989. Cooking up referring expressions. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, Vancouver, British Columbia, Canada. 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.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, 36(5):799–836.

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

Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the virtual to the real world: Referring to objects in real-world spatial scenes. In *Proceedings of EMNLP 2015*. Association for Computational Linguistics.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

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)*, pages 787–798, Doha, Qatar.

Jungjun Kim, Hanbin Ko, and Jialin Wu. 2020. CoNAN: A complementary neighboring-based attention network for referring expression generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1952–1962, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Emiel Krahmer and Kees van Deemter. 2019. *Computational Generation of Referring Expressions: An Updated Survey*.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

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 *European conference on computer vision*, pages 740–755. Springer.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring expression generation and comprehension via attributes. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Jingyu Liu, Wei Wang, Liang Wang, and Ming-Hsuan Yang. 2020. Attribute-guided attention for referring expression generation and comprehension. *IEEE Transactions on Image Processing*, 29:5244–5258.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

R. Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134.

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

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.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, pages 95–104. Association for Computational Linguistics.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *CoRR*, abs/1907.09358.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2020. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech & Language*, 68:101184.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Simeon Schüz, Ting Han, and Sina Zarrieß. 2021. Diversity as a by-product: Goal-oriented language generation leads to linguistic variation. In *Proceedings of the 22th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 175–185. Association for Computational Linguistics.

M. Tanaka, Takayuki Itamochi, K. Narioka, Ikuro Sato, Y. Ushiku, and T. Harada. 2019. Generating easy-to-understand referring expressions for target identifications. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5793–5802.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

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

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

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. 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)*, pages 610–620, Berlin, Germany. Association for Computational Linguistics.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.