

Visual Referring Expression Recognition: What Do Systems Actually Learn?

Volkan Cirik, Louis-Philippe Morency, Taylor Berg-Kirkpatrick
Carnegie Mellon University
{vcirik, morency, tberg}@cs.cmu.edu

Abstract

We present an empirical analysis of state-of-the-art systems for referring expression recognition – the task of identifying the object in an image referred to by a natural language expression – with the goal of gaining insight into how these systems reason about language and vision. Surprisingly, we find strong evidence that even sophisticated and linguistically-motivated models for this task may ignore linguistic structure, instead relying on shallow correlations introduced by unintended biases in the data selection and annotation process. For example, we show that a system trained and tested on the input image *without the input referring expression* can achieve a precision of 71.2% in top-2 predictions. Furthermore, a system that predicts only the *object category* given the input can achieve a precision of 84.2% in top-2 predictions. These surprisingly positive results for what should be deficient prediction scenarios suggest that careful analysis of what our models are learning – and further, how our data is constructed – is critical as we seek to make substantive progress on grounded language tasks.

1 Introduction

There has been increasing interest in modeling natural language in the context of a visual grounding. Several benchmark datasets have recently been introduced for describing a visual scene with natural language (Chen et al., 2015), describing or localizing specific objects in a scene (Kazemzadeh et al., 2014; Mao et al., 2016), answering natural language questions about the scenes (Antol et al., 2015), and performing visually grounded dialogue (Das et al., 2016). Here, we focus on referring expression recognition (RER) – the task of identifying the object in an image that is referred to by a natural language expression produced by a human (Kazemzadeh et al., 2014; Mao et al., 2016; Hu

et al., 2016; Rohrbach et al., 2016; Yu et al., 2016; Nagaraja et al., 2016; Hu et al., 2017).

Recent work on RER has sought to make progress by introducing models that are better capable of reasoning about linguistic structure (Hu et al., 2017; Nagaraja et al., 2016) – however, since most of the state-of-the-arts systems involve complex neural parameterizations, what these models actually learn has been difficult to interpret. This is concerning because several post-hoc analyses of related tasks (Zhou et al., 2015; Devlin et al., 2015; Agrawal et al., 2016; Jabri et al., 2016; Goyal et al., 2016) have revealed that some positive results are actually driven by superficial biases in datasets or shallow correlations without deeper visual or linguistic understanding. Evidently, it is hard to be completely sure if a model is performing well for the right reasons.

To increase our understanding of how RER systems function, we present several analyses inspired by approaches that probe systems with perturbed inputs (Jia and Liang, 2017) and employ simple models to exploit and reveal biases in datasets (Chen et al., 2016a). First, we investigate whether systems that were designed to incorporate linguistic structure actually require it and make use of it. To test this, we perform perturbation experiments on the input referring expressions. Surprisingly, we find that models are robust to shuffling the word order and limiting the word categories to nouns and adjectives. Second, we attempt to reveal shallower correlations that systems might instead be leveraging to do well on this task. We build two simple systems called Neural Sieves: one that completely *ignores* the input referring expression and another that only predicts the category of the referred object from the input expression. Again, surprisingly, both sieves are able to identify the correct object with surprising precision in top-2 and top-3 predictions. When these two simple systems are com-

bined, the resulting system achieves precisions of 84.2% and 95.3% for top-2 and top-3 predictions, respectively. These results suggest that to make meaningful progress on grounded language tasks, we need to pay careful attention to what and how our models are learning, and whether our datasets contain exploitable bias.

2 Related Work

Referring expression recognition and generation is a well studied problem in intelligent user interfaces (Chai et al., 2004), human-robot interaction (Fang et al., 2012; Chai et al., 2014; Williams et al., 2016), and situated dialogue (Kennington and Schlangen, 2017). Kazemzadeh et al. (2014) and Mao et al. (2016) introduce two benchmark datasets for referring expression recognition. Several models that leverage linguistic structure have been proposed. Nagaraja et al. (2016) propose a model where target and supporting objects (i.e. objects that are mentioned in order to disambiguate the target object) are identified and scored jointly. The resulting model is able to localize supporting objects without direct supervision. Hu et al. (2017) introduce a compositional approach for the RER task. They assume that the referring expression can be decomposed into a triplet consisting of the target object, the supporting object, and their spatial relationship. This structured model achieves state-of-the-art accuracy on the Google-Ref dataset. Cirik et al. (2018) propose a type of neural modular network (Andreas et al., 2016) where the computation graph is defined in terms of a constituency parse of the input referring expression.

Previous studies on other tasks have found that state-of-the-art systems may be successful for reasons different than originally assumed. For example, Chen et al. (2016b) show that a simple logistic regression baseline with carefully defined features can achieve competitive results for reading comprehension on CNN/Daily Mail datasets (Hermann et al., 2015), indicating that more sophisticated models may be learning relatively simple correlations. Similarly, Gururangan et al. (2018) reveal bias in a dataset for semantic inference by demonstrating a simple model that achieves competitive results *without looking at the premise*.

3 Analysis by Perturbation

In this section, we would like to analyze how the state-of-the-art referring expression recognition systems utilize linguistic structure. We conduct

experiments with perturbed referring expressions where various aspects of linguistic structure are obscured. We perform three types of analyses: the first one studying syntactic structure (Section 3.2), the second focusing on the importance of word categories (Section 3.3), and the final one analyzing potential biases in the dataset (Section 3.4).

3.1 Analysis Methodology

To perform our analysis, we take two state-of-the-art systems CNN+LSTM-MIL (Nagaraja et al., 2016) and CMN (Hu et al., 2017) and train them from scratch with perturbed referring expressions. We note that the perturbation experiments explained in next subsections are performed on all train and test instances. All experiments are done on the standard train/test splits for the Google-Ref dataset (Mao et al., 2016). Systems are evaluated using the precision@ k metric, the fraction of test instances for which the target object is contained in the model’s top- k predictions. We provide further details of our experimental methodology in Section 4.1.

3.2 Syntactic Analysis by Permuting Word Order

In English, word order is important for correctly understanding the syntactic structure of a sentence. Both models we analyze use Recurrent Neural Networks (RNN) (Elman, 1990) with Long Short-Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997). Previous studies have shown that recurrent architectures can perform well on tasks where word order and syntax are important: for example, tagging (Lample et al., 2016), parsing (Sutskever et al., 2014), and machine translation (Bahdanau et al., 2014). We seek to determine whether recurrent models for RER depend on syntactic structure.

Premise 1: *Randomly permuting the word order of an English referring expression will obscure its syntactic structure.*

We train CMN and CNN+LSTM-MIL with shuffled referring expressions as input and evaluate their performance.

| Model | No Perturbation | Shuffled | Δ |
|--------------|-----------------|----------|----------|
| CMN | .705 | .675 | -.030 |
| LSTM+CNN-MIL | .684 | .630 | -.054 |

Table 1: Results for Shuffling Word Order for Referring Expressions. Δ is the difference between no perturbation and shuffled version of the same system.

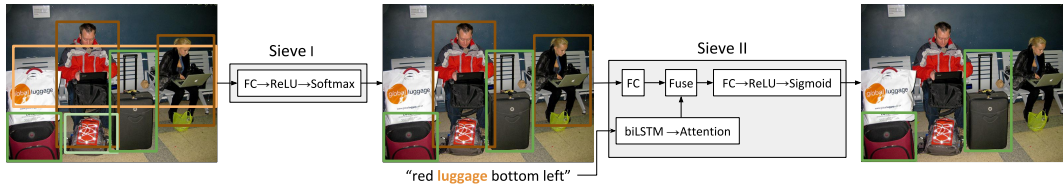


Figure 1: Overview of Neural Sieves. Sieve I filters object types having multiple instances. Sieve II filters objects of one category mentioned in referring expression. Objects of the same category have the same color frames. Best seen in color.

Table 1 shows accuracies for models with and without shuffled referring expressions. The column with Δ shows the difference in accuracy compared to the best performing model without shuffling. The drop in accuracy is surprisingly low. Thus, we conclude that these models do not stongly depend on the syntactic structure of the input expression and may instead leverage other, shallower, correlations.

3.3 Lexical Analysis by Discarding Words

Following the analysis presented in Section 3.2, we are curious to study what other aspects of the input referring expression may be essential for state-of-the-art performance. If syntactic structure is largely unimportant, it may be that spatial relationships can be ignored. Spatial relationships between objects are usually represented by prepositional phrases and verb phrases. In contrast, simple descriptors (e.g. green) and object types (e.g. table) are most often represented by adjectives and nouns, respectively. By discarding all words in the input that are not nouns or adjectives, we hope to test whether spatial relationships are actually important to state-of-the-art models. Notably, both systems we test were specifically designed to model object relationships.

Premise 2: *Keeping only nouns and adjectives from the input expression will obscure the relationships between objects that the referring expression describes.*

Table 2 shows accuracies resulting from training and testing these models on only the nouns and adjectives in the input expression. Our first observation is that the accuracies of models drop the most when we discard the nouns (the rightmost column in Table 2). This is reasonable since nouns

| Models | Noun & Adj (Δ) | Noun (Δ) | Adj (Δ) |
|--------------|-------------------------|-------------------|------------------|
| CMN | .687 (-.018) | .642 (-.063) | .585 (-.120) |
| LSTM+CNN-MIL | .644 (-.040) | .597 (-.087) | .533 (-.151) |

Table 2: Results with discarded word categories. Numbers in parentheses are Δ , the difference between the best performing version of the original model.

define the types of the objects referred to in the expression. Without nouns, it is extremely difficult to identify which objects are being described. Second, although both systems we analyze model the relationship between objects, discarding verbs and prepositions, which are essential in determining the relationship among objects, does not drastically effect their performance (the second column in Table 2). This may indicate the superior performance of these systems does not specifically come from their modeling approach for object relationships.

3.4 Bias Analysis by Discarding Referring Expressions

Goyal et al. (2016) show that some language and vision datasets have exploitable biases. Could there be a dataset bias that is exploited by the models for RER?

Premise 3: *Discarding the referring expression entirely and keeping only the input image creates a deficient prediction problem: achieving high-performance on this task indicates dataset bias.*

We train CMN by removing all referring expressions from train and test. We call this model “image-only” since it ignores the referring expression and will only use the input image. We compare the CMN “image-only” model with the state-of-the-art configuration of CMN and a *random baseline*. Table 3 shows precision@ k results. The “image-

| Model | P@1 | P@2 | P@3 | P@4 | P@5 |
|------------------|------|------|------|------|------|
| CMN | .705 | .926 | .979 | .993 | .998 |
| CMN “image-only” | .411 | .731 | .885 | .948 | .977 |
| Random Baseline | .204 | .403 | .557 | .669 | .750 |

Table 3: Results with discarded referring expressions. Surprisingly, the top-2 prediction (73.1%) of the “image-only” model is better than the top prediction of the state-of-the-art (70.5%).

only” model is able to surpass the random baseline by a large margin. This result indicates that the dataset is biased, likely as a result of the data selection and annotation process. During the construction of the dataset, Mao et al. (2016) annotate an object box only if there are at least 2 to 4 objects

of the same type in the image. Thus, only a subset of object categories ever appear as targets because some object types rarely occur multiple times in an image. In fact, out of 90 object categories in MSCOCO, 43 of the object categories are selected as target objects less than 1% of the time they occur in images. This potentially explains the relative high performance of the “image-only” system.

3.5 Discussion

The previous analyses indicate that exploiting bias in the data selection process and leveraging shallow linguistic correlations with the input expression may go a long way towards achieving high performance on this dataset. First, it may be possible to simplify the decision of picking an object to a much smaller set of candidates without even considering the referring expression. Second, because removing all words except for nouns and adjectives only marginally hurt performance for the systems tested, it may be possible to further reduce the set of candidates by focusing only on simple properties like the category of the target object rather than its relations with the environment or with adjacent objects.

4 Neural Sieves

We introduce a simple pipeline of neural networks, Neural Sieves, that attempt to reduce the set of candidate objects down to a much smaller set that still contains the target object given an image, a set of objects, and the referring expression describing one of the objects.

Sieve I: Filtering Unlikely Objects. Inspired by the results from Section 3.4, we design an “image-only” model as the first sieve for filtering *unlikely* objects. For example in Figure 1, Sieve I filters out the backpack and the bench from the list of bounding boxes since there is only one instance of these object types. We use a similar parameterization of one of the baselines (CMN_{LOC}) proposed by Hu et al. (2017) for Sieve I and train it by only providing spatial and visual features for the boxes, ignoring the referring expression. More specifically, for visual features r^{vis} of a bounding boxes of an object, we use Faster-RCNN (Ren et al., 2015). We use 5-dimensional vectors for spatial features $r^{spat} = [\frac{x_{min}}{W_V}, \frac{y_{min}}{H_V}, \frac{x_{max}}{W_V}, \frac{y_{max}}{W_V}, \frac{A_r}{A_V}]$ where A_r is the size and $[x_{min}, y_{min}, x_{max}, y_{max}]$ are coordinates for bounding box r and A_V, W_V, H_V are the area, the width, and the height of the input image

V . These two representations are concatenated as $r^{vis,spat} = [r^{vis}, r^{spat}]$ for a bounding box r .

We parameterize Sieve I with a list of bounding boxes R as the input with parameter set Θ_I as follows:

$$s_I = W_I^{score} r^{vis,spat} \quad (1)$$

$$f_I(R; \Theta_I) = softmax(s_I) \quad (2)$$

Each bounding box is scored using a matrix W_I^{score} . Scores for all bounding boxes are then fed to softmax to get a probability distribution over boxes. The learned parameter Θ_I is the scoring matrix W_I^{score} .

Sieve II: Filtering Based on Objects Categories

After filtering *unlikely* objects based only on the image, the second step is to determine which object category to keep as a candidate for prediction, filtering out the other categories. For instance, in Figure 1, only instances of suitcases are left as candidates after determining which type of object the input expression is talking about. To perform this step, Sieve II takes the list of object candidates from Sieve I and keeps objects having the same object category as the referred object. Unlike Sieve I, Sieve II uses the referring expression to filter bounding boxes of objects. We again use the baseline model of CMN_{LOC} from the previous work (Hu et al., 2017) for the parametrization of Sieve II with a minor modification: instead of predicting the referred object, we make a binary decision for each box of whether the object in the box is the same category as the target object.

More specifically, we parameterize Sieve II as follows:

$$\hat{r}^{vis,spat} = W_{II}^{vis,spat} r^{vis,spat} \quad (3)$$

$$z_{II} = \hat{r}^{vis,spat} \odot f_{att}(T) \quad (4)$$

$$\hat{z}_{II} = z_{II} / \|z_{II}\|_2 \quad (5)$$

$$s_{II} = W_{II}^{score} \hat{z}_{s2} \quad (6)$$

$$f_{II}(T, R; \Theta_{II}) = sigmoid(s_{II}) \quad (7)$$

We encode the referring expression T into an embedding with $f_{att}(T)$ which uses an attention mechanism (Bahdanau et al., 2014) on top of a 2-layer bidirectional LSTM (Schuster and Paliwal, 1997).

We project bounding box features $r^{vis,spat}$ to the same dimension as the embedding of referring expression (Eq 3). Text and box representations are element-wise multiplied to get z_{II} as a joint representation of the text and bounding box (Eq 4). We L2-normalize to produce \hat{z}_{II} (Eq 5, 6). Box scores

| Model | precision@ k | Accuracy |
|---------------------|----------------|----------|
| CMN | 1 | .705 |
| CMN | 2 | .926 |
| CMN | 3 | .979 |
| LSTM+CNN-MIL | 1 | .684 |
| LSTM+CNN-MIL | 2 | .907 |
| LSTM+CNN-MIL | 3 | .972 |
| Neural Sieve I | 1 | .401 |
| Neural Sieve I | 2 | .712 |
| Neural Sieve I | 3 | .866 |
| Neural Sieve I + II | 1 | .488 |
| Neural Sieve I + II | 2 | .842 |
| Neural Sieve I + II | 3 | .953 |

Table 4: Precision@ k accuracies for Neural Sieves and state-of-the-art systems. Note that even without using the referring expression, Sieve I is able to reduce the number of candidate boxes to 3 for 86.6% of the instances. When we further predict the type of objects with Sieve II, the number of candidate boxes is reduced to 2 for 84.2% of the instances.

s_{II} are calculated with a linear projection of the joint representation (Eq 6) and fed to the sigmoid function for a binary prediction for each box. The learned parameters Θ_{II} are $W_{II}^{vis,spat}$, W_{II}^{score} , and parameters of the encoding module f_{att} .

4.1 Filtering Experiments

We are interested in determining how accurate these simple neural sieves can be. High accuracy here would give a possible explanation for the high performance of more complex models.

Dataset. For our experiments, we use Google-Ref (Mao et al., 2016) which is one of the standard benchmarks for referring expression recognition. It consists of around 26K images with 104K annotations. We use their Ground-Truth evaluation setup where the ground truth bounding box annotations from MSCOCO (Lin et al., 2014) are provided to the system as a part of the input. We used the split provided by Nagaraja et al. (2016) where splits have disjoint sets of images. We use precision@ k for evaluating the performance of models.

Implementation Details. To train our models, we used stochastic gradient descent for 6 epochs with an initial learning rate of 0.01 and multiplied by 0.4 after each epoch. Word embeddings were initialized using GloVe (Pennington et al., 2014) and finetuned during training. We extracted features for bounding boxes using the fc7 layer output of Faster-RCNN VGG-16 network (Ren et al., 2015) pre-trained on MSCOCO dataset (Lin et al., 2014). Hyperparameters such as hidden layer size of LSTM networks were picked based on the best

validation score. For perturbation experiments, we did not perform any grid search for hyperparameters. We used hyperparameters of the previously reported best performing model in the literature. We released our code for public use¹.

Baseline Models. We compare Neural Sieves to the state-of-the-art models from the literature. **LSTM + CNN - MIL** Nagaraja et al. (2016) score *target object-context object* pairs using LSTMs for processing the referring expression and CNN features for bounding boxes. The pair with the highest score is predicted as the referred object. They use Multi-Instance Learning for training the model. **CMN** (Hu et al., 2017) is a neural module network with a tuple of object-relationship-subject nodes. The text encoding of tuples is calculated with a two-layer bi-directional LSTM and an attention mechanism (Bahdanau et al., 2014) over the referring expression.

4.2 Results

Table 4 shows the precision scores. The referred object is in the top-2 candidates selected by Sieve I 71.2% of the time and in the top-3 predictions 86.6% of the time. Combining both sieves into a pipeline, these numbers further increase to 84.2% for top-2 predictions and to 95.3% for top-3 predictions. Considering the simplicity of Neural Sieve approach, these are surprising results: two simple neural network systems, the first one ignoring the referring expression, the second predicting only object type, are able to reduce the number of candidate boxes down to 2 on 84.2% of instances.

5 Conclusion

We have analyzed two RER systems by variously perturbing aspects of the input referring expressions: shuffling, removing word categories, and finally, by removing the referring expression entirely. Based on this analysis, we proposed a pipeline of simple neural sieves that captures many of the easy correlations in the standard dataset. Our results suggest that careful analysis is important both while constructing new datasets and while constructing new models for grounded language tasks. The techniques used here may be applied more generally to other tasks to give better insight into what our models are learning and whether our datasets contain exploitable bias.

¹<https://github.com/volkancirik/neural-sieves-refexp>

References

- Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1955–1960. <https://doi.org/10.18653/v1/D16-1203>.
- Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Joyce Y Chai, Pengyu Hong, and Michelle X Zhou. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, pages 70–77.
- Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, pages 33–40.
- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016a. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2358–2367. <https://doi.org/10.18653/v1/P16-1223>.
- Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016b. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2358–2367. <http://www.aclweb.org/anthology/P16-1223>.
- 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*.
- Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Phillippe Morency. 2018. Using syntax to ground referring expressions in natural images. In *32nd AAAI Conference on Artificial Intelligence (AAAI-18)*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *arXiv preprint arXiv:1611.08669*.
- Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.
- Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14(2):179–211.
- Rui Fang, Changsong Liu, and Joyce Yue Chai. 2012. Integrating word acquisition and referential grounding towards physical world interaction. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, pages 109–116.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *arXiv preprint arXiv:1612.00837*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.
- Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks.
- Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.
- Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *European conference on computer vision*. Springer, pages 727–739.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.
- Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language* 41:43–67.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.
- 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*. Springer, pages 740–755.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 11–20.
- Varun Nagaraja, Vlad Morariu, and Larry Davis. 2016. Modeling context between objects for referring expression understanding. In *ECCV*.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. pages 91–99.
- Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*. Springer, pages 817–834.
- Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.
- Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated open world reference resolution for human-robot dialogue. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, pages 311–318.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, pages 69–85.
- Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2015. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.