# Is this a Child, a Girl or a Car? Exploring the Contribution of Distributional Similarity to Learning Referential Word Meanings

**Sina Zarrieß** and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{sina.zarriess,david.schlangen}@uni-bielefeld.de

## Abstract

There has recently been a lot of work trying to use images of referents of words for improving vector space meaning representations derived from text. We investigate the opposite direction, as it were, trying to improve visual word predictors that identify objects in images, by exploiting distributional similarity information during training. We show that for certain words (such as entry-level nouns or hypernyms), we can indeed learn better referential word meanings by taking into account their semantic similarity to other words. For other words, there is no or even a detrimental effect, compared to a learning setup that presents even semantically related objects as negative instances.

## 1 Introduction

Someone who knows the meaning of the word *child* will most probably know a) how to distinguish children from other entities in the real world and b) that *child* is related to other words, such as *girl*, *boy*, *mother*, etc. Traditionally, these two aspects of lexical meaning—which, following (Marconi, 1997), we may call *referential* and *inferential*, respectively—have been modeled in quite distinct settings. Semantic similarity has been a primary concern for distributional models of word meaning that treat words as vectors which are aggregated over their contexts, cf. (Turney and Pantel, 2010; Erk, 2016). Identifying visual referents of words, on the other hand, is a core requirement for verbal human/robot interfaces (HRI) (Roy et al., 2002; Tellex et al., 2011; Matuszek et al., 2012; Krishnamurthy and Kollar, 2013; Kennington and Schlangen, 2015). Here, word meanings have been modeled as predictors that can be applied to the visual representation of an object and predict referential appropriateness for that object.

This paper extends upon recent work on learning models of referential word use on large-scale corpora of images paired with referring expressions (Schlangen et al., 2016). As in previous approaches in HRI, that work treats words during training and application as independent predictors, with no relations between them. Our starting assumption here is that this misses potentially useful information: e.g., that the costs for confusing referents of *child* vs. *boy* should be much lower than for confusing referents of *child* vs. *car*. We thus investigate whether knowledge about semantic similarities between words can be exploited to learn more accurate visual word predictors, accounting for this intuition that certain visual object distinctions are semantically more important or costly than others.

We explore two methods for informing visual word predictors about semantic similarities in a distributional space: a) by sampling negative instances of word such that they contain more dissimilar objects, b) by labeling instances with a more fine-grained real-valued supervision signal derived from pairwise distributional similarities between object names. We find that the latter, similarity-based training method leads to substantial improvements for particular words such as entry-level nouns or hypernyms, whereas predictors for other words such as adjectives do not benefit from distributional knowledge. These results suggest that, in principle, semantic relatedness might be promising knowledge source for training more accurate visual models of referential word use, but it also supports recent findings showing that distributional models do not capture all aspects of semantic relatedness equally well (Rubinstein et al., 2015; Nguyen et al., 2016).



*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 86–91,
Valencia, Spain, April 3-7, 2017. 

## 2 Models for Referential Word Meaning

We model referential word meanings as predictors that can be applied to the visual representation of an object and return a score indicating the appropriateness of the word for denoting the object. We describe now different ways of defining these predictors with respect to semantic similarity.

**Words as Predictors (WAP)** We train a binary classifier for each word $w$ in the vocabulary. The training set for each word $w$ is built as follows: all visual objects in an "image + referring expression" corpus that have been referred to as $w$ are used as positive instances, the remaining objects as negative instances. Thus, the set of object images divides into $w$ and $\neg w$, with the consequence that all negative instances are considered equally dissimilar from $w$. The classifiers are trained with logistic regression (using $\ell 1$ penalty). (This is the (Schlangen et al., 2016) model.)

**Undersampling similar objects (WAP-NOSIM)** As discussed above, it is intuitive to assume that a visual classifier that distinguishes referents of a word from other objects in an image should be less penalized for making errors on objects that are categorically related. For instance, the classifier for *child* should be less penalized for giving high probabilities to referents of *boy* than to referents of *car*. A straightforward way to introduce these differences during training is by undersampling negative instances that have been referred to by very similar words. (E.g., undersampling *boy* instances as negative instances for the *child* classifier.) This should allow the word classifier to focus on visual distinctions between objects that are semantically more important. When compiling the training set of a WAP-NOSIM classifier for word $w$, we look at its 10 most similar words in the vocabulary according to a distributional model (trained with word2vec, see below) and remove their instances from the set of negative instances $\neg w$.

**Word as Similarity Predictors (SIM-WAP)** Instead of removing similar objects from the training set of a word model, we can task the model with directly learning similarities, by training it as a linear regression on a continuous output space. When building the training set for such a word predictor $w$, instead of simply dividing objects into $w$ and $\neg w$ instances, we label each object with a real-valued similarity obtained from cosine similarity between $w$ and $v$ in a distributional vector space, where $v$ is the word used to refer to the object. Object instances where $v = w$ (i.e., the positive instances in the binary setup) have maximal similarity; the remaining instances have a lower value which is more or less close to maximal similarity. This then yields a more fine-grained labeling of what is uniformly considered as negative instances in the binary set-up.

We transform the cosine similarities between words in our vocabulary into standardised $z$ scores (mean: 0, sd: 1). When there are several word candidates used for an object in the corpus, we simply use the word $v$ that has maximal similarity to our target word $w$. The predictors are trained with Ridge Regression.

## 3 Experimental Set-up

We focus on assessing to what extent similarity-based visual word predictors capture the referential meaning of a word in a more accurate way, and distinguish its potential referents from other random objects. To factor out effects of compositionality and context that arise in reference generation or resolution, we measure how well a predictor for a word $w$ is able to retrieve from a sampled test set objects that have been referred to by $w$ (Schlangen et al., 2016; Zarrieß and Schlangen, 2016a) evaluate on full referring expressions).

**Data** As training data, we use the training split of the REFERIT corpus collected by (Kazemzadeh et al., 2014), which is based on the medium-sized SAIAPR image collection (Grubinger et al., 2006) (99.5k image regions). For testing, we use the training section of REFCOCO corpus collected by (Yu et al., 2016), which is based on the MSCOCO collection (Lin et al., 2014) containing over 300k images with object segmentations. This gives us a large enough test set to make stable predictions about the quality of individual word predictors, which often only have a few positive instances in the test set of the REFERIT corpus. We follow (Schlangen et al., 2016) and select words with a minimum frequency of 40 in these two data sets, which gives us a vocabulary of 793 words.

**Evaluation** For each word, we sample a test set that includes all its positive instances, and positive vs. negative instances at a ratio of 1:100. We apply the word classifier to all test instances and assess how well it identifies (retrieves) its posi-

| | | Avg. Precision | |
|---|---|---|---|
| | | referit | refcoco |
| Vocab | *# samples (avg.)* | *1055* | *8176* |
| | WAP | **0.369** | 0.183 |
| | WAP-NOSIM | 0.358 | 0.179 |
| | SIM-WAP | 0.354 | **0.188** |
| Entry-level Nouns | *# samples (avg.)* | *2143* | *11275* |
| | WAP | **0.506** | 0.228 |
| | WAP-NOSIM | 0.497 | 0.211 |
| | SIM-WAP | 0.489 | **0.296** |

Table 1: Mean average precision for word predictors, on small (referit) and large (refcoco) test set

| | Avg. Prec. | | | | |
|---|---|---|---|---|---|
| word | WAP | SIM-WAP | #train | #test | most similar to |
| animal | 0.45 | 0.60 | 37 | 533 | animals, dog, cat |
| animals | 0.31 | 0.53 | 9 | 13 | animal, birds, sheep |
| plant | 0.41 | 0.68 | 41 | 123 | plants, shrubs, flower |
| plants | 0.58 | 0.82 | 18 | 17 | plant, shrubs, flowers |
| bird | 0.58 | 0.76 | 45 | 196 | birds, parrot, turtle |
| birds | 0.06 | 0.22 | 11 | 7 | bird, animals, parrot |
| vehicle | 0.44 | 0.67 | 9 | 101 | car, cars, truck |
| food | 0.21 | 0.44 | 13 | 669 | meat, drink, eating |

Table 2: Evaluation of word predictors for hypernyms in singular and plural on REFCOCO

tive instances, i.e. visual objects that have been referred to by the word. We measure this using average precision, corresponding to the area under the curve (AUC) metric. In Section 4, we report performance over the entire vocabulary and the subset of entry-level nouns extracted from annotations in the REFERIT corpus (Kazemzadeh et al., 2014).

**Image and Word Embeddings** Following (Schlangen et al., 2016), we derive representations of our visual inputs with a convolutional neural network, "GoogLeNet" (Szegedy et al., 2015), that was trained on data from the ImageNet corpus (Deng et al., 2009), and extract the final fully-connected layer before the classification layer, to give us a 1024 dimensional representation of the region. We add 7 features that encode information about the region relative to the image: the (relative) coordinates of two corners, its (relative) area, distance to the center, and orientation of the image. The full representation hence is a vector of 1031 features. As distributional word vectors, we use the `word2vec` representations provided by Baroni et al. (2014) (trained with 5-word context window, 10 negative samples, 400 dimensions).

## 4 Results

**Overall** In Table 1, we show the means of the average precision scores achieved by the individual word predictors. Generally, the differences between the overall means for the different models are mostly small, but we will see below that there are more pronounced differences when looking at particular parts of the vocabulary. On the REFERIT test set, the simple binary classifiers (WAP) have a slight advantage over the similarity-based methods. On REFCOCO, SIM-WAP performs best, improving slightly over wac on the entire vocabulary and substantially when looking at the subset of entry-level nouns. By contrast, the WAP-NOSIM classifiers (trained with under sampling of similar objects) perform slightly worse as compared to the standard binary classifiers on all test sets. First, this suggests that there is an effect of corpus or domain. Performance is substantially lower on REFCOCO than on REFERIT, but the similarity-based predictors generalize better across the data sets. Second, this shows that under sampling is not a good way of dealing with similar objects when training word predictors whereas in similarity-based training the model does take advantage of distributional knowledge, at least in certain cases.

**Individual Words** As shown in Table 1, the similarity-based training has a strong positive effect for entry-level nouns, whereas the effect on the overall vocabulary is rather small. This further suggests that distributional similarities improve certain word predictors substantially, whereas others might be affected even negatively. Therefore, in the following, we report average precision for individual words, namely for those cases where similarity-based regression has the strongest positive or negative effect as compared to binary classification (see Tables 3 and 4 showing average precision scores, number of positive instances of the word in the train and test set, and their semantic neighbours in the vocabulary, according to the vector space). We also look at hypernyms (Table 2) which are not easy to learn in realistic referring expression data as more specific nouns are usually more common or natural (Ordonez et al., 2016).

**Where similarities help** Table 3 shows results for words where SIM-WAP improves most over the binary WAP model on REFCOCO. It seems that especially some low-frequent words benefit from knowledge about object similarities, improving their average precision by more than 30% or 40% on the test set that contains more positive instances even than were observed during training.

| word | AP WAP | AP SIM-WAP | # train | # test | most similar to |
|---|---|---|---|---|---|
| *# positive training instances* $< 50$ | | | | | |
| trailer | 0.16 | 0.54 | 1 | 28 | truck, vehicle, car |
| suv | 0.42 | 0.79 | 2 | 40 | vehicle, car, cars |
| pillow | 0.21 | 0.57 | 2 | 66 | pillows, bed, nightstand |
| doors | 0.10 | 0.44 | 6 | 11 | door, curtains, window |
| sheep | 0.40 | 0.74 | 1 | 524 | lamb, goat, animals |
| *# positive training instances* $> 50$ | | | | | |
| kid | 0.22 | 0.43 | 74 | 1641 | kids, boy, girl |
| boy | 0.22 | 0.41 | 55 | 1330 | girl, boys, kid |
| bike | 0.50 | 0.69 | 76 | 842 | bicycle, motorcycle, car |
| horse | 0.57 | 0.73 | 55 | 757 | dog, donkey, cow |
| bottle | 0.39 | 0.55 | 61 | 213 | bottles, jar, glasses |

Table 3: Top 5 improvements for SIM-WAP over WAP, for rare and more-frequent words

| word | AP wac | AP SIM-WAP | #train | #test | most similar to |
|---|---|---|---|---|---|
| *# positive training instances* $< 50$ | | | | | |
| pie | 0.44 | 0.10 | 1 | 86 | cake, cheese, pastry |
| surf | 0.56 | 0.20 | 1 | 43 | surfboard, snowboard |
| number | 0.44 | 0.07 | 1 | 172 | four, two, three |
| anywhere | 0.59 | 0.21 | 88 | 34 | anything, anyone |
| monitor | 0.65 | 0.15 | 2 | 228 | watch, handle, laptop |
| *# positive training instances* $> 50$ | | | | | |
| pink | 0.18 | 0.10 | 52 | 814 | purple, blue, yellow |
| green | 0.19 | 0.11 | 257 | 1393 | blue, yellow, greens |
| area | 0.17 | 0.09 | 167 | 253 | city, land, square |
| big | 0.15 | 0.06 | 74 | 737 | huge, bigger, biggest |
| upper | 0.25 | 0.07 | 116 | 633 | lower, middle |

Table 4: Top 5 degradations for SIM-WAP over WAP, shown for rare and frequent words

Similarly, predictors for hypernyms and their plural versions improve substantially, see Table 2. All of these example words have semantic neighbours that are also visually similar. Similarity-based training of word predictors hence is very beneficial for rare words (during training) that have near-synonymy relations to other words in the corpus. The positive effect here probably relates to "feature-sharing", as the predictor for "trailer" is allowed to learn from the positive instances of "truck", rather than having to discriminate between the referents of the two words.

**Where similarities do not help** In Table 4, we can see results for words where similarity-based training does not help. For words with more than 50 training instances, distributional similarities degrade performance most for adjectives and words expressing visual attributes (color, shape, location). In these cases, distributional similarities group attributes from the same scale (color or location), but do not account for the fact that these are visually distinct, such as in the case of e.g 'upper' and 'lower'. Similarly, distributional similarities between colors seem to be misleading rather than helpful, cf. (Zarrieß and Schlangen, 2016b) for a study on color adjectives on the same corpus. This effect seems to be related to findings on antonyms in distributional modeling (Nguyen et al., 2016). Overall, as words corresponding to attributes are quite frequent in the referring expression data, the negative effect of similarity-based training seems to balance out the positive effect found for certain nouns in the overall evaluation. Similar effects can also be found for nouns where semantic similarities predicted by a distributional model seem to diverge strongly from visual similarity that would be helpful for learning the referential meaning of the word, e.g. 'monitor' and 'watch'.

## 5 Discussion and Conclusion

Even with access to powerful state-of-the-art object recognizers that classify objects in images into thousands of categories with high accuracy, it is still a challenging task to model referential meanings of individual words and to capture various visual distinctions between semantically similar and dissimilar words and their referents. In contrast to abstract objects labels that are annotated consistently in image corpora, word use in referring expressions is more flexible, and subject to a range of communicative factors, in such a way that e.g. some instances of *child* will be named not by this but by similar words.

Our findings suggest that linking distributional similarity to models for visual word predictors capturing referential meaning is promising to account for the fact that the negative instances used for training word predictors vary in their degree of semantic similarity to the positive instances of a word. We explored two different ways of integrating this information—by undersampling and by directly predicting similarity—and found the prediction approach to work better, especially for low- and medium-frequent words that have a range of lexically similar neighbors in the model's vocabulary.

In a similar vein, zero-shot learning approaches to object recognition (Frome et al., 2013; Lazaridou et al., 2014; Norouzi et al., 2013) have transferred visual knowledge from known object classes to unknown classes via distributional similarity. Here, we show that visual knowledge can be

transferred between words in a corpus of referring expressions, by taking into account their semantic relation while learning.

Our results suggest that the exploration of joint improvement of inferential (i.e., similarity-based) and referential aspects of meaning should be a fruitful avenue for future work.

## Acknowledgments

We acknowledge support by the Cluster of Excellence "Cognitive Interaction Technology" (CITEC; EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

## References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63, April.

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 13–23, Genoa, Italy.

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.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China, July. Association for Computational Linguistics.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland, June. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C.Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision ECCV 2014*, volume 8693, pages 740–755. Springer International Publishing.

Diego Marconi. 1997. *Lexical Competence*. MIT Press, Cambridge, Mass., USA.

Cynthia Matuszek, Nicholas Fitzgerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the International Conference on Machine Learning (ICML 2012)*.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 454–459, Berlin, Germany. Association for Computational Linguistics.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. 2013. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations (ICLR)*.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2016. Learning to name objects. *Communications of the ACM*, 59(3):108–115, February.

Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. 2002. A trainable spoken language understanding system for visual object selection. In *Proceedings of the International Conference on Speech*

*and Language Processing 2002 (ICSLP 2002)*, Colorado, USA.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730, Beijing, China, July. Association for Computational Linguistics.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany, August. Association for Computational Linguistics.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. In *AAAI Conference on Artificial Intelligence*, pages 1507–1514.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 69–85. Springer International Publishing, Cham.

Sina Zarrieß and David Schlangen. 2016a. 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, August. Association for Computational Linguistics.

Sina Zarrieß and David Schlangen. 2016b. Towards Generating Colour Terms for Referents in Photographs: Prefer the Expected or the Unexpected? In *Proceedings of the 9th International Natural Language Generation conference*, pages 246—-255. Association for Computational Linguistics.