Obtaining referential word meanings from visual and distributional information: Experiments on object naming

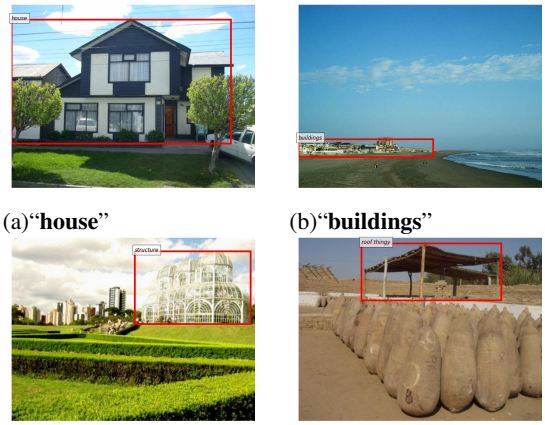
Sina Zarrieß and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{sina.zarriess,david.schlangen}@uni-bielefeld.de

Abstract

We investigate object naming, which is an important sub-task of referring expression generation on real-world images. As opposed to mutually exclusive labels used in object recognition, object names are more flexible, subject to communicative preferences and semantically related to each other. Therefore, we investigate models of referential word meaning that link visual to lexical information which we assume to be given through distributional word embeddings. We present a model that learns individual predictors for object names that link visual and distributional aspects of word meaning during training. We show that this is particularly beneficial for zero-shot learning, as compared to projecting visual objects directly into the distributional space. In a standard object naming task, we find that different ways of combining lexical and visual information achieve very similar performance, though experiments on model combination suggest that they capture complementary aspects of referential meaning.

1 Introduction

Expressions referring to objects in visual scenes typically include a word naming the type of the object: E.g., *house* in Figure 1 (a), or, as a very general type, *thingy* in Figure 1 (d). Determining such a name is a crucial step for referring expression generation (REG) systems, as many other decisions concerning, e.g., the selection of attributes follow from it (Dale and Reiter, 1995; Krahmer and Van Deemter, 2012). For a long time, however, research on REG mostly assumed the availability of symbolic representations of referent and scene, and sidestepped questions about how speakers actually choose these names, due to the lack of models capable of capturing what a word like *house* refers to in the real world.



(a)"**house**" (b)"**buildings**"

(c)"large **structure**" (d)"roof **thingy**"

Figure 1: Examples of object names in the REFERIT corpus referring to instances of buildings

Recent advances in image processing promise to fill this gap, with state-of-the-art computer vision systems being able to classify images into thousands of different categories (e.g. Szegedy et al. (2015)). However, classification is not naming (Ordonez et al., 2016). Standard object classification schemes are inherently "flat", and treat object labels as mutually exclusive (Deng et al., 2014). A state-of-the-art object recognition system would be trained to classify the object in e.g. Figure 1 (a) as either *house* or *building*, ignoring the lexical similarity between these two names. In contrast, humans seem to be more flexible as to the chosen level of generality. Depending on the prototypicality of the object to name, and possibly other visual properties, a general name might be more or less appropriate. For instance, a robin can be named *bird*, but a penguin is better referred

to as "*penguin*" (Rosch, 1978); along the same lines, the rather unusual building in Figure 1 (c) that is not easy to otherwise categorise was named "*structure*".

Other work at the intersection of image and language processing has investigated models that learn to directly associate visual objects with a continuous representation of word meaning, i.e. through cross-modal transfer into distributional vector spaces (Frome et al., 2013; Norouzi et al., 2013). Here, the idea is to exploit a powerful model of lexical similarity induced from large amounts text for being able to capture inherent lexical relations between object categories. Thus, under the assumption that such semantic spaces represent, in some form at least, taxonomic knowledge, this makes labels on different levels of specificity available for a given object. Moreover, if the mapping is sufficiently general, it should be able to map objects to an appropriate label, even if during training of the mapping this label has not been seen (*zero-shot learning*).

While cross-modal transfer seems to be a conceptually attractive model for learning object names, it is based on an important assumption that, in our view, has not received sufficient attention in previous works: it assumes that a given distributional vector space constitutes an optimal target representation that visual instances of objects can be mapped to. However, distributional representations of word meaning are known to capture a rather fuzzy notion of lexical similarity, e.g. *car* is similar to *van* and to *street*. A cross-modal transfer model is "forced" to learn to map objects into the same area in the semantic space if their names are distributionally similar, but regardless of their actual visual similarity. Indeed, we have found in a recent study that the contribution of distributional information to learning referential word meanings is restricted to certain types of words and does not generalize across the vocabulary (Zarrieß and Schlangen, 2017).

The goal of this work is to learn a model of referential word meaning that makes accurate object naming predictions and goes beyond treating words as independent, mutually exclusive labels in a flat classification scheme. We extend upon work on learning models of referential word use from corpora of images paired with referring expressions (Schlangen et al., 2016; Zarrieß and Schlangen, 2017) that treats words as individual predictors capturing referential appropriateness. We explore different ways of linking these predictors to distributional knowledge, during application and during training. We find that these different models achieve very similar performance in a standard object naming task, though experiments on model combination suggest that they capture complementary aspects of referential meaning. In a zero-shot setup of an object naming task, we find that combining lexical and visual information during training is most beneficial, outperforming variants of cross-modal transfer.

2 Related Work

Grounding and Reference An early example for work in REG that goes beyond Dale and Reiter (1995)'s dominant symbolic paradigm is Deb Roy's work from the early 2000s (Roy et al., 2002; Roy, 2002, 2005). Roy et al. (2002) use computer vision techniques to process a video feed, and to compute colour, positional and spatial features. These features are then associated in a learning process with certain words, resulting in an association of colour features with colour words, spatial features with prepositions, etc., and based on this, these words can be interpreted with reference to the scene currently presented to the video feed. Whereas Roy's work still looked at relatively simple scenes with graphical objects, research on REG has recently started to investigate set-ups based on real-world images (Kazemzadeh et al., 2014; Gkatzia et al., 2015; Zarrieß and Schlangen, 2016; Mao et al., 2015). Importantly, the low-level visual features that can be extracted from these scenes correspond less directly to particular word classes. Moreover, the visual scenes contain many different types of objects, which poses new challenges for REG. For instance, Zarrieß and Schlangen (2016) find that semantic errors related to mismatches between nouns (e.g. the system generates *tree* vs. *man*) are particularly disturbing for users. Whereas Zarrieß and Schlangen (2016) propose a strategy to avoid object names when the systems confidence is low, we focus on improving the generation of object names, using distributional knowledge as an additional source. Similarly, Ordonez et al. (2016) have studied the problem of deriving appropriate object names, or so-called entry-level categories, from the output of an object recognizer. Their approach focusses on linking abstract object categories in ImageNet

to actual words via various translation procedures. We are interested in learning referential appropriateness and extensional word meanings directly from actual human referring expressions (REs) paired with objects in images, using an existing object recognizer for feature extraction.

Multi-modal distributional semantics Distributional semantic models are a well-known method for capturing lexical word meaning in a variety of tasks (Turney and Pantel, 2010; Mikolov et al., 2013; Erk, 2016). Recent work on multimodal distributional vector spaces (Feng and Lapata, 2010; Silberer and Lapata, 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015b; Kottur et al., 2016) has aimed at capturing semantic similarity even more accurately by integrating distributional and perceptual features associated with words (mostly taken from images) into a single representation.

Cross-modal transfer Rather than fusing different modalities into a single, joint space, other work has looked at cross-modal mapping between spaces. Herbelot and Vecchi (2015) present a model that learns to map vectors in a distributional space to vectors in a set-theoretic space, showing that there is a functional relationship between distributional information and conceptual knowledge representing quantifiers and predicates. More related to our work are cross-modal mapping models,that learn to transfer from a representation of an object or image in the visual space to a vector in a distributional space (Socher et al., 2013; Frome et al., 2013; Norouzi et al., 2013; Lazaridou et al., 2014). Here, the motivation is to exploit the rich lexical knowledge encoded in a distributional space for learning visual classifications. In practice, these models are mostly used for zero-shot learning where the test set contains object categories not observed during training. When tested on standard object recognition tasks, transfer, however, comes at a price. Frome et al. (2013) and Norouzi et al. (2013) both find that it slightly degrades performance as compared to a plain object classification using standard accuracy metrics (called flat "hit @k metric" in their paper). Interestingly though, Frome et al. (2013) report better performance using "hierarchical precision", which essentially means that transfer predicts words that are ontologically closer to the gold label and makes "semantically more reasonable errors". To the best of our knowledge, this pattern has not been systematically investigated any further. Another known problem with cross-modal transfer is that it seems to generalize less well than expected, i.e. tends to reproduce word vectors observed during training (Lazaridou et al., 2015a). In this work, we present a model that exploits distributional knowledge for learning referential word meaning as well, but explore and compare different ways of combining visual and lexical aspects of referential word meaning.

3 Task and Data

We define *object naming* as follows: Given an object x in an image, the task is to predict a word w that could be used as the head noun of a realistic referring expression. (Cf. discussion above: "*bird*" when naming a robin, but "*penguin*" when naming a penguin.) To get at this, we develop our approach using a corpus of referring expressions produced by human users under natural, interactive conditions (Kazemzadeh et al., 2014), and train and test on the corresponding head nouns in these REs. This is similar to picture naming setups used in psycholinguistic research (cf. Levelt et al. (1991)) and based on the simplifying assumption that the name used for referring to an object can be determined successfully without looking at other objects in the image.

We now summarise the details of our setup:

Corpus We train and test on the REFERIT corpus (Kazemzadeh et al., 2014), which is based on the SAIAPR image collection (Grubinger et al., 2006) (99.5k image regions;120K REs). 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.

Names For most of our experiments, we only use a subset of this vocabulary, namely the set of object names. As the REs contain nouns that cannot be considered to be object names (*background, bottom*, etc.), we extract a list of names from the semantically annotated held-out set released with the REFERIT. These correspond to 'entry-level' nouns mentioned in Kazemzadeh et al. (2014). This gives us a list of 159 names. This set corresponds to the majority of object names in the corpus: out of the 99.5K available image regions, we use 80K for training and testing. Thus, our experiments are on a smaller scale as compared

to (Ordonez et al., 2016). Nevertheless, the data is challenging, as the corpus contains references to objects that fall outside of the object labeling scheme that available object recognition systems are typically optimized for, cf. Hu et al. (2015)'s discussion on "stuff" entities such "sky" or "grass" in the REFERIT data. For testing, we remove relational REs (containing a relational preposition such as 'left of X'), because here we cannot be sure that the head noun of the target is fully informative; we also remove REs with more than one head noun from our list (i.e. these are mostly relational expressions as well such as 'girl laughing at boy'). We pair each image region from the test set with its corresponding names from the remaining REs.

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), which was trained on 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, thus representing each object as a vector of 1031 features. As distributional word vectors, we use the `word2vec` representations provided by Baroni et al. (2014) (trained with CBOW, 5-word context window, 10 negative samples, 400 dimensions).

4 Three Models of Interfacing Visual and Distributional Information

4.1 Direct Cross-Modal Mapping

Following Lazaridou et al. (2014), referential meaning can be represented as a translation function that projects visual representations of objects to linguistic representations of words in a distributional vector space. Thus, in contrast to standard object recognition systems or the other models we will use here, cross-modal mapping does not treat words as individual labels or classifiers, but learns to directly predict continuous representations of words in a vector space, such as the space defined by the `word2vec` embeddings that we use in this work. This model will be called TRANSFER below.

During training, we pair each object with the distributional embedding of its name, and use standard Ridge regression for learning the transformation. Lazaridou et al. (2014) and Lazaridou et al. (2015a) test a range of technical tweaks and different algorithms for cross-modal mapping. For ease of comparison with other models, we stick with simple Ridge Regression in this work.

For decoding, we map an object into the distributional space, and retrieve the nearest neighbors of the predicted vector using cosine similarity. In theory, the model should generalize easily to words that it has not observed in a pair with an object during training as it can map an object anywhere in the distributional space.

4.2 Lexical Mapping Through Individual Word Classifiers

Another approach is to keep visual and distributional information separate, by training a separate visual classifier for each word w in the vocabulary. Predictions can then be mapped into distributional space during application time via the vectors of the predicted words. Here, we use Schlangen et al. (2016)'s WAC model, building the training set for each word w as follows: all visual objects in a corpus that have been referred to as w are used as positive instances, the remaining objects as negative instances. Thus, the classifiers learn to predict referential appropriateness for individual words based on the visual features of the objects they refer to, in isolation of other words.

During decoding, we apply all word classifiers from the model's vocabulary to the given object, and take the argmax over the individual word probabilities. The model predicts names directly, without links into a distributional space.

In order to extend the model's vocabulary for zero-shot learning, we follow Norouzi et al. (2013) and associate the top n words with their corresponding distributional vector and compute the convex combination of these vectors. Then, in parallel to cross-modal mapping, we retrieve the nearest neighbors of the combined embedding from the distributional space. Thus, with this model, we use two different modes of decoding: one that projects into distributional space, one that only applies the available word classifiers.

We did some small-scale experiments to find an optimal value for n, similar to Norouzi et al. (2013). In our case, performance started to decrease systematically with $n > 10$, but did not differ significantly for values below 10. In Section 5, we will report results for n set to 5 and 10.

4.3 Word Prediction via Cross-Modal Similarity Mapping

Finally, we implement an approach that combines ideas from cross-modal mapping with the WAC model: we train individual predictors for each word in the vocabulary, but, during training, we exploit lexical similarity relations encoded in a distributional space. Instead of treating a word as a binary classifier, we annotate its training instances with a fine-grained similarity signal according to their object names. 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 that was used to refer to the object. Thus, we task the model with jointly learning similarities and referential appropriateness, by training it with Ridge regression on a continuous output space. 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 is the SIM-WAP model, recently proposed in Zarrieß and Schlangen (2017).

Importantly, and going beyond Zarrieß and Schlangen (2017), this model allows for an innovative treatment of words that only exist in a distributional space (without being paired with visual referents in the image corpus): as the predictors are trained on a continuous output space, no genuine positive instances of a word's referent are needed. When training a predictor for such a word w, we use all available objects from our corpus and annotate them with the expected lexical similarity between w and the actual object names v, which for all objects will be below the maximal value that marks genuine positive instances. During decoding, this model does not need to project its predictions into a distributional space, but it simply applies all available predictors to the object, and takes the argmax over the predicted referential appropriateness scores.

5 Experiment 1: Naming Objects

This Section reports on experiments in a standard setup of the object naming task where all object names are paired with visual instances of their referents during training. In a comparable task, i.e. object recognition with known object categories, cross-modal projection or transfer approaches have been reported to perform worse than standard object classification methods (Frome et al., 2013; Norouzi et al., 2013). This seems to suggest that lexical or at least distributional knowledge is detrimental when learning what a word refers to in the real world and that referential meaning should potentially be learned from visual object representation only.

5.1 Model comparison

Setup We use the train/test split of REFERIT data as in (Schlangen et al., 2016). We consider image regions with non-relational referring expressions that contain at least one of the 159 head nouns from the list of entry-level nouns (see section 3). This amounts to 6208 image regions for testing and 73K instances for training.

Results Table 1 shows accuracies in the object naming task for the TRANSFER, WAC and SIM-WAP models according to their accuracies in the top n, including two variants of WAC where its top 5 and top 10 predictions are projected into the distributional space. Overall, the models achieve very similar performance. However, there is an interesting pattern when comparing accuracies @1 and @2 to accuracies in the top 5 predictions. Thus, looking at accuracies for the top (two) predictions, the various models that link referential meaning to word representations in the distributional space all perform slightly worse than the plain WAC model, i.e. individual word classifiers trained on visual features only. This might suggest that certain aspects of referential word meaning are learned less accurately when mapping from visual to distributional space (which replicates results reported in the literature on standard object recognition benchmarks). On the other hand, the SIM-WAP model is on a par with WAC in terms of the @5 accuracy. This effect suggests that distributional knowledge that SIM-WAP has access to during training sometimes distracts the model from predicting the exact name chosen by a human speaker, but that SIM-WAP is still able to rank it among the most probable names. As a simple accuracy-based evaluation is not suited to fully explain this pattern, we carry out a more detailed analysis in Section 5.3.

	hit @k(%)		
	@1	@2	@5
transfer	48.34	60.49	74.89
wac	**49.34**	**61.86**	**75.35**
wac, project top5	48.73	61.10	74.07
wac, project top10	48.68	61.23	74.31
sim-wap	48.13	60.60	**75.40**

Table 1: Accuracies in object naming

	hit @k(%)		
	1	5	10
sim-wap + transfer	49.10	61.78	75.81
sim-wap + wac	51.10	63.45	77.92
transfer + wac	51.13	63.76	77.84
wac + transfer + sim-wap	**52.19**	**64.71**	**78.40**

Table 2: Object naming acc., combined models

5.2 Model combination

In order to get more insight into why the TRANSFER and SIM-WAP models produce slightly worse results than individual visual word classifiers, we now test to what extent the different models are complementary and combine them by aggregating over their naming predictions. If the models are complementary, their combination should lead to more confident and accurate naming decisions.

Setup We combine TRANSFER, SIM-WAP and WAC by aggregating the scores they predict for different object names for a given object. During testing, we apply all models to an image region and consider words ranked among the top 10. We first normalize the referential appropriateness scores in each top-10 list and then compute their sum. This aggregation scheme will give more weight to words that appear in the top 10 list of different models, and less weight to words that only get top-ranked by a single model. We test on the same data as in Section 5.1.

Results Table 2 shows that all model combinations improve over the results of their isolated models in Table 1, suggesting that WAC, TRANSFER and SIM-WAP indeed do capture complementary aspects of referential word meaning. On their own, the distributionally informed models are less tuned to specific word occurrences than the visual word classifiers in the WAC model, but they can add useful information which leads to a clear overall improvement. We take this as a promising finding, supporting our initial hypothesis that knowledge on lexical distributional meaning should and can be exploited when learning how to use words for reference.

	Av. cosine similarity			
	among top k		gold - top k	
	5	10	5	10
transfer	**0.32**	**0.27**	**0.28**	**0.25**
wac	0.18	0.20	0.18	0.16
sim-wap	**0.32**	0.26	**0.28**	**0.25**

Table 3: Cosine similarities between word2vec embeddings of nouns generated in the top k

5.3 Analysis

Figure 2 illustrates objects from our test set where the combination of TRANSFER, SIM-WAP and WAC predicts an accurate name, whereas the models in isolation do not. These examples give some interesting insight into why the models capture different aspects of referential word meaning.

Word Similarities Many of the examples in Figure 2 suggest that the object names ranked among the top 3 by the TRANSFER and SIM-WAP model are semantically similar to each other, whereas WAC generates object names on top that describe very different underlying object categories, such as *seal* / *rock* in Figure 2(a), *animal* / *lamp* in Figure 2(g) or *chair* / *shirt* in Figure 2(c). To quantify this general impression, Table 3 shows cosine similarities among words in the top n generated by our models, using their word2vec embeddings. The average cosine similarity between words in our vocabulary is 0.17. The TRANSFER and SIM-WAP model rank words on top that are clearly more similar to each other than word pairs on average, whereas words ranked top by the WAC model are more dissimilar to each other. Another remarkable finding is that the words generated by TRANSFER and SIM-WAP are not only more similar among the top predictions, but also more similar to the gold name (Table 3 , right columns). This result is noteworthy since the accuracies for the top predictions shown in Table 1 are slightly below WAC. In general, this suggests that there is a trade-off between optimizing a model of referential word meaning to exact naming decisions, or tailoring it to make lexically consistent predictions. This parallels findings by Frome et al. (2013) who found that their transfer-based object recognition made "semantically more reasonable" errors than a standard convolutional network while

not improving accuracies for object recognition, see discussion in Section 2. Additional evaluation metrics, such as success rates in a human evaluation (cf. Zarrieß and Schlangen (2016)), would be an interesting direction for more detailed investigation here.

Word Use But even though the WAC classifiers lack knowledge on lexical similarities, they seem to able to detect relatively specific instances of word use such as *hut* in Figure 2(b), *shirt* in 2(c) or *lamp* in 2(h). Here, the combination with TRANSFER and SIM-WAP is helpful to give more weight to the object name that is taxonomically correct (sometimes pushing up words below the top-3 and hence not shown in Figure 2). In Figure 1(e), SIM-WAP and TRANSFER give more weight to typical names for persons, whereas WAC top-ranks more unusual names, reflecting that the person is difficult to identify visually. Another observation is that the mapping models have difficulties dealing with object names in singular and plural. As these words have very similar representations in the distributional space, they are often predicted as likely variants among the top 10 by SIM-WAP and TRANSFER, whereas the WAC model seems to predict inappropriate plural words less often among the top 3. Such specific phenomena at the intersection of visual and semantic similarity have found very little attention in the literature. We will investigate them further in our Experiments on zero-shot naming in the following Section.

6 Zero-Shot Naming

Zero-shot learning is an attractive prospect for REG from images, as it promises to overcome dependence on pairings of visual instances and natural names being available for all names, if visual/referential data can be generalised from other types of information. Previous work has looked at the feasibility of zero-shot learning as a function of semantic similarity or ontological closeness between unknown and known categories, and confirmed the intuition that the task is harder the less close unknown categories are to known ones (Frome et al., 2013; Norouzi et al., 2013).

Our experiments on object naming in Section 5 suggest that lexical similarities encoded in a distributional space might not always fully carry over to referential meaning. This could constitute an additional challenge for zero-shot learning, as distributional similarities might be misleading when the model has to fully rely on them for learning referential word meanings. Therefore, the following experiments investigate the performance of our models in zero-shot naming as a function of the lexical relation between unknown and known object names, i.e. namely hypernyms and singular/plurals. Both relations are typically captured by distributional models of word meaning in terms of closeness in the vector space, but their visual and referential relation is clearly different.

6.1 Vocabulary Splits and Testsets

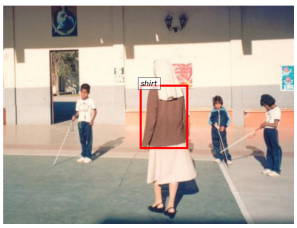
Random As in previous work on zero-shot learning, we consider zero-shot naming for words of varying degrees of similarity. We randomly split our 159 names from Experiment 1 into 10 subsets. We train the models on 90% of the nouns (and all their visual instances in the image corpus) and test on the set of image regions that are named with words which the model did not observe during training. Results reported in Table 4 on the random test set correspond to averaged scores from cross-validation over the 10 splits.

Hypernyms We manually split the model's vocabulary into set of hypernyms (see Appendix A) and the remaining nouns. We train the models on those 84K image regions that where not named with a hypernym, and test on 8895 image regions that were named with a hypernym in the corpus. We checked that for each of these hypernyms, the vocabulary contains at least one or two names that can be considered as hyponyms, i.e. the model sees objects during training that are instances of *vehicle* for example, but never encounters actual uses of that name. This test set is particularly interesting from an REG perspective, as objects named with very general terms by human speakers are often difficult to describe with more common, but more specific terms, as is illustrated by the uses of *structure* and *thingy* in Figure 1.

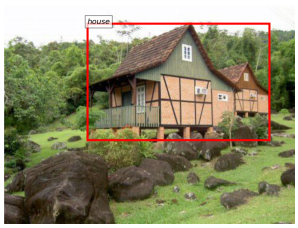
Singulars/Plurals We pick 68 words from our vocabulary that can be grouped into 34 singular-plural noun pairs (see Appendix A). From each pair, we randomly include the singular or plural noun in the set of zero-shot nouns. Thus, we make sure that the model encounters singular and plural names during training, but it never encounters both variants of a name. This results training split of 23K image regions and a test split of 13825 instances.

(a)
wac: seal, rock, water
sim-wap: side, rock,rocks
transfer: rocks, rock, water
combination: rock

(b)
wac: cactus, hut, mountain
sim-wap: side, rock, mountain
transfer: mountain, rocks, rock
combination: hut



(c)
wac: chair, shirt, guy
sim-wap: woman, man, girl
transfer: door, woman, windov
combination: shirt



(d)
wac: roof, house, building
sim-wap: building, house, trees
transfer: building, house, trees
combination: house

(e)
wac: chick, person, guy
sim-wap: man, person, woman
transfer: man, guy, girl
combination: person

(f)
wac: bush, bushes, tree
sim-wap: trees, tree, grass
transfer: trees, tree, bushes
combination: bushes

(g)
wac: animal, lamp, table
sim-wap: man, girl, person
transfer: man, clouds, cloud
combination: person

(h)
wac: post, light, lamp
sim-wap: tree, sky, pole
transfer: tree, sky, trees
combination: lamp

Figure 2: Examples from object naming experiment where model combination is accurate

Zero-shot names	Model	full vocab @1	full vocab @2	full vocab @5	full vocab @10	disjoint vocab @1	disjoint vocab @2
Random	transfer	0.05	2.38	16.57	35.71	41.49	62.34
	wac, project top10	0.00	4.42	21.16	39.17	38.03	58.07
	wac, project top5	0.00	4.39	21.63	40.01	37.46	57.36
	sim-wap	**3.71**	**13.13**	**36.49**	**54.44**	**42.28**	**64.26**
Hypernyms	transfer	0.07	1.25	7.75	29.93	**59.88**	**73.88**
	wac, project top10	0.00	3.01	15.55	36.99	50.51	66.33
	wac, project top5	0.00	2.78	16.75	38.13	47.73	64.38
	sim-wap	**3.16**	**10.33**	**31.14**	**49.62**	57.55	70.15
Singulars/Plurals	transfer	0.01	22.84	44.30	72.85	34.56	51.79
	wac, project top10	0.00	22.21	43.43	68.95	31.46	48.76
	wac, project top5	0.00	22.18	43.93	69.33	31.46	48.88
	sim-wap	**15.39**	**34.73**	**56.62**	**77.32**	**37.24**	**54.02**

Table 4: Accuracies in zero-shot object naming on different vocabulary splits

6.2 Evaluation

Some previous work on zero-shot image labeling assumes additional components that first identify whether an image should be labelled by a known or unknown word (Frome et al., 2013). We follow Lazaridou et al. (2014) and let the model decide whether to refer to an object by a known or unknown name. Related to that, distinct evaluation procedures have been used in the literature on zero-shot learning:

Testing on full vocabulary A realistic way to test zero-shot learning performance is to consider all words from a given vocabulary during testing, though the testset only contains instances of objects that have been named with a 'zero-shot word' (for which no visual instances were seen during training). Accuracies in this setup reflect how well the model is able to generalize, i.e. how often it decides to deviate from the words it was trained on, and (implicitly) predicts that the given object requires a "new" name. In case of the (i) hypernym and (ii) singular/plural test set, this accuracy also reflects to what extent the model is able to detect cases where (i) a more general or vague term is needed, where (ii) an unknown singular/plural counterpart of a known object type occurs.

Testing on disjoint vocabulary Alternatively, the model's vocabulary can be restricted during testing to zero-shot words only, such that names encountered during training and testing are disjoint, see e.g. (Lampert et al., 2009, 2013). This setup factors out the generalization problem, and assesses to what extent a model is able to capture the referential meaning of a word that does not have instances in the training data.

6.3 Results

As compared to Experiment 1 where models achieved similar performance, differences are more pronounced in the zero-shot setup, as shown in Table 4. In particular, we find that the SIM-WAP model which induces individual predictors for words that have not been observed in the training data is clearly more successful than TRANSFER or WAC that project predictions into the distributional space. When tested on the full vocabulary, we find that TRANSFER and WAC very rarely generate names whose referents were excluded from training, which is in line with observations made by Lazaridou et al. (2015a). The SIM-WAP predictors generalize much better, in particular on the singular/plural testset.

An interesting exception is the good performance of the TRANSFER model on the hypernym test set, when evaluated with a disjoint vocabulary. This corroborates evidence from Experiment 1, namely that the transfer model captures taxonomic aspects of object names better than the other models. Projection via individual word classifiers, on the other hand, seems to generalize better than TRANSFER, at least when looking at accuracies @2 ... @10. Thus, combining several vectors predicted by a model of referential word meaning can provide additional information, as compared to mapping an object to a single vector in distributional space. More work is needed to establish how these approaches can be integrated more effectively.

7 Discussion and Conclusion

In this paper, we have investigated models of referential word meaning, using different ways of combining visual information about a word's referent and distributional knowledge about its lexical similarities. Previous cross-modal mapping models essentially force semantically similar objects to be mapped into the same area in the semantic space regardless of their actual visual similarity. We found that cross-modal mapping produces semantically appropriate and mutually highly similar object names in its top-n list, but does not preserve differences in referential word use (e.g. appropriatness of *person* vs. *woman*) especially within the same semantic field. We have shown that it is beneficial for performance in standard and zero-shot object naming to treat words as individual predictors that capture referential appropriateness and are only indirectly linked to a distributional space, either through lexical mapping during application or through cross-modal similarity mapping during training. As we have tested these approaches on a rather small vocabulary, which may limit generality of conclusions, future work will be devoted to scaling up these findings to larger test sets, as e.g. recently collected through conversational agents (Das et al., 2016) that circumvent the need for human-human interaction data. Also from an REG perspective, various extensions of this approach are possible, such as the inclusion of contextual information during object naming and its combination with attribute selection.

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). We thank the anonymous reviewers for their very valuable, very detailed and highly interesting comments.

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)*. Association for Computational Linguistics, Baltimore, Maryland, pages 238–247. http://www.aclweb.org/anthology/P14-1023.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *CoRR* abs/1611.08669. http://arxiv.org/abs/1611.08669.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*. Springer, pages 48–64.

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

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics* 9(17):1–63. https://doi.org/10.3765/sp.9.17.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 91–99.

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*, Curran Associates, Inc., pages 2121–2129.

Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the virtual to the realworld: Referring to objects in real-world spatial scenes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1936–1942. http://aclweb.org/anthology/D15-1224.

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)*. Genoa, Italy, pages 13–23.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 22–32. http://aclweb.org/anthology/D15-1003.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2015. Natural language object retrieval. *CoRR* abs/1511.04164. http://arxiv.org/abs/1511.04164.

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)*. Doha, Qatar, pages 787–798.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 36–45. http://www.aclweb.org/anthology/D14-1005.

Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 4985–4994.

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

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*. IEEE, pages 951–958.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(3):453–465.

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.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015a. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. 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)*. Association for Computational Linguistics, Beijing, China, pages 270–280. http://www.aclweb.org/anthology/P15-1027.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 153–163. http://www.aclweb.org/anthology/N15-1016.

Willem JM Levelt, Herbert Schriefers, Dirk Vorberg, Antje S Meyer, Thomas Pechmann, and Jaap Havinga. 1991. The time course of lexical access in speech production: A study of picture naming. *Psychological review* 98(1):122.

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. http://arxiv.org/abs/1511.02283.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, NIPS'13, pages 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959.

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. *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. *Commun. ACM* 59(3):108–115.

Eleanor Rosch. 1978. Principles of Categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, Lawrence Erlbaum, Hillsdale, N.J., USA, pages 27—-48.

Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends in Cognitive Sciene* 9(8):389–396.

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.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language* 16(3).

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 721–732. http://www.aclweb.org/anthology/P14-1068.

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.

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.

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.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 610–620. http://www.aclweb.org/anthology/P16-1058.

Sina Zarrieß and David Schlangen. 2017. Is this a child, a girl or a car? exploring the contribution of distributional similarity to learning referential word meanings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 86–91. http://aclweb.org/anthology/E17-2014.

A Vocabulary Splits for Zero-Shot Naming

Hypernyms animal, animals, plant, plants, vehicle, person, persons, food, thing, object, area, things, thingy, toy, anyone, clothes, dish, building, land, structure, item, water

Singulars/Plurals . . .

. . . training on instances of: animals, plants, cars, people, buildings, trees, man, kid, guy, girl, boy, flower, bird, hill, orange, cloud, curtain, window, shrub, apple, light, house, glass, bottle, dude, leg, book, wall, bananas, carrots, pillows, bushes, mountains, bags

. . . testing on instances of: animal, plant, car, person, building, tree, men, kids, guys, girls, boys, flowers, birds, hills, oranges, clouds, curtains, windows, shrubs, apples, lights, houses, glasses, bottles, dudes, legs, books, walls, banana, carrot, pillow, bush, mountain, bag