Deriving continous grounded meaning representations from referentially structured multimodal contexts

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

Corpora of referring expressions paired with their visual referents are a good source for learning word meanings directly grounded in visual representations. Here, we explore additional ways of extracting from them word representations linked to multi-modal context: through expressions that refer to the *same* object, and through expressions that refer to *different* objects in the *same scene*. We show that continuous meaning representations derived from these contexts capture complementary aspects of similarity, even if not outperforming textual embeddings trained on very large amounts of raw text when tested on standard similarity benchmarks. We propose a new task for evaluating grounded meaning representations—detection of potentially co-referential phrases—and show that it requires precise denotational representations of attribute meanings, which our method provides.

1 Introduction

Various routes for linking language to extra-linguistic context have been explored in recent years. A lot of research has looked at integrating visual representations, either directly (Matuszek et al., 2012; Krishnamurthy and Kollar, 2013; Yu et al., 2016; Schlangen et al., 2016) or through mapping into a multi-modal distributional space (Feng and Lapata, 2010; Bruni et al., 2012; Kiela and Bottou, 2014; Lazaridou et al., 2015). Young et al. (2014) have explored a less direct link, by representing the extension of phrasal expressions as sets of images, and deriving from this a precise notion of denotational similarity. In very recent work, Cocos and Callison-Burch (2017) use spatial context from geo-located tweets to induce word embeddings that capture situational similarity between lexical items.

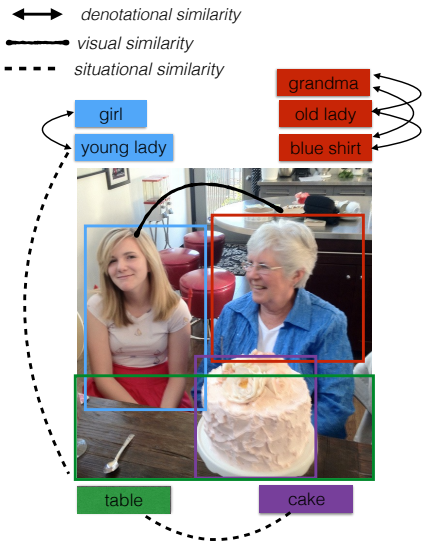

Figure 1: Dimensions of context in referential, visually grounded language, and similarity relations that can be derived from it, image from MSCOCO (Lin et al., 2014))

In this paper, we explore an approach that combines aspects of several of these paths. Starting point is the observation that corpora of exophoric referring expressions provide richly structured contexts that go beyond just linking individual expressions with their denotations. As an example consider the scene in Figure 1 depicting several referents and corresponding referring expressions produced by different speakers. This scene provides a learner not only with an example of a referent for the word *lady*, it also provides the information that *lady* can co-refer with *girl*, and that its denotations can spatially / situationally co-occur with those of *table* and *cake*. From these types of information we infer word embeddings, following the method from Levy and Goldberg (2014) for training embeddings on arbi-

trary non-linear context, and we show that these capture complementary aspects of word similarity that purely textual induction methods conflate. We also show that these representations handle a more directly referential similarity task better.

2 Word Embeddings from Multi-Modal Referential Contexts

We base our study on the REFERIT and REFCOCO corpus (Kazemzadeh et al., 2014; Yu et al., 2016) building upon image collections by (Grubinger et al., 2006) and (Lin et al., 2014); for the latter, we also use referring expressions collected by Mao et al. (2015). This corpus gives us visual scenes containing sets of objects, $s = o_1, \ldots, o_n$. Each object is associated with a set of referring expressions $r_1, \ldots, r_m$; and we use a standard method (a ConvNet) for providing a visual representation vis_i for it. Each referring expression, in turn, is defined as a linear sequence of words $r_i = w_1 \ldots w_k$. In the following, we structure this context into four dimensions—visual, textual, situational and denotational—which we use to derive different word embeddings.

2.1 Textual Context (TXT)

We learn standard distributional word embeddings from our corpus, ignoring extra-linguistic context. We train a skip-gram model (Mikolov et al., 2013) with negative sampling with window width 5, 300 dimensions. For comparison, we also use the textual word embeddings provided by Baroni et al. (2014), trained on a much larger web corpus (5-word context window, 10 negative samples, 400 dimensions). We distinguish the two textual embeddings using the subscripts TXT_{ref}, TXT_{web}.

2.2 Visual Grounding (VIS)

Given a set of referring expressions containing the word w and their corresponding referent $(o_j, r_j), w \in r_j$, we can derive a visual context for the word w by averaging over the visual representations of its referents vis_j, as proposed for instance by Kiela and Bottou (2014). The visual context of a word can be seen as a 'visual prototype'. 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. Following (Schlangen et al., 2016), we also add 7 features that encode information about the region relative to the image, the full representation hence is a vector of 1031 features. Each word is then represented as the average over its visual vectors.

2.3 Situational Grounding (SIT)

We also train word embeddings (dim. 300) that predict words paired with their situational context, following the method by Levy and Goldberg (2014). This captures similarities between words occurring for different objects in the same scene, e.g. *cake* in the context of *table* in Figure 1. Given a pair of referring expressions $(r_i, o_i), (r_j, o_j), o_i \neq o_j$, r_i and r_j are co-situational expressions. Thus, for a word $w_i \in r_i$, we consider all words $w_j \in r_j$ as its situational context. In practice, we compute situational contexts only for the head nouns of each referring expression, as we expect situational similarities to be useful for capturing similarities between nouns.

2.4 Denotational Grounding (DEN)

As our data typically records multiple co-referential expressions for an object (3 expressions on average in the REFCOCO data), we define the denotational context based on sets of expressions referring to the same object $(r_1, o_i) \ldots (r_n, o_i)$. For a word $w_i \in r_i$, we consider all words w_{j_l} (with $w_{j_l} \in r_j$) as denotational context, where r_j and r_i refer to the same object. When two words occur in a denotational context, we have strong evidence that they are semantically compatible, i.e. can refer to the same objects as *girl* and *lady* in Figure 1 do. Similar to our training procedure for situational embeddings, we now learn 300-dimensional word embeddings that predict occurrences of a word based on co-referential contexts, pairing each word with all words from referring expressions describing the same object.

3 Word Similarity and Relatedness

We now have four different continuous representations for words; in the following, we evaluate them for how well they predict semantic relations.

Similarity We evaluate on some similarity data sets, reporting Spearman ρ correlations between human ratings and cosine similarities for word vectors. We use the MEN (Bruni et al., 2012) and

Silberer and Lapata (2014)'s data with semantic (SemSim) and visual similarity (VisSim) ratings.

Compatibility As generic semantic similarity judgements are known to be "fuzzy" (Faruqui et al., 2016), we also evaluate on Kruszewski and Baroni (2015)'s benchmark on semantic compatibility. They define two words as being *semantically compatible* "if they can potentially refer to the same thing". We expect our denotational and visual embeddings to be highly useful for this task. We report unsupervised results obtained from cosine similarities between word embeddings.

Hypernym Directionality We adopt an evaluation procedure by Kiela et al. (2015b) on hypernym pairs in the BLESS data set (Baroni and Lenci, 2011). Given a general (e.g. 'animal') and a concrete noun (e.g. 'dog') that stand in the hypernym relation, the task is to identify the noun that is more general. Lazaridou et al. (2015) found that the generality or concreteness of a noun's meaning is reflected in the entropy of its embedding, and we adopt that measure for our purposes. Thus, we compute entropies of our word embeddings and report accuracies corresponding to the proportion of noun pairs where the entropy of the more general noun is higher than the more concrete noun.

Vocabulary We intersect the vocabularies covered by the different embeddings, which amounts to 1960 words in total. We restrict evaluation to the corresponding word pairs in the above data sets, coverage is reported in Table 1.

Results As shown in Table 1, the performance of embeddings learned on referring expression corpora are generally below state-of-the-art distributional vectors trained on large web corpora. However, some interesting tendencies can be observed by comparing embeddings learned from different context dimensions. Denotational embeddings in isolation provide a precise representation of meaning that outperforms the other types of embeddings on semantic similarity judgements in MEN and SemSim, and detects hypernym directionality most accurately. An interesting exception is the compatibility data set where visual embeddings clearly outperform textual and denotational embeddings. Situational embeddings perform less well than textual and denotational embeddings but, interestingly, are similar in performance to visual embeddings on semantic similarity, suggesting that visual and situational similarity seem to be equally important aspects of general semantic similarity. Concatenation of denotational and textual embeddings yields the best results for correlations with human similarity judgements. This is expected as denotational similarity is probably too restricted for generic semantic similarity. We experimented with further embedding combinations, but only the fusion of the textual and denotational dimension outperformed the embeddings obtained from a particular grounding dimension.

Model	MEN	SemSim	VisSim	Compat.	Hyp.Dir.
# pairs	*989*	*2041*	*2041*	*4843*	*334*
VIS	0.404	0.469	0.427	**0.241**	78.14
TXT_{ref}	0.550	0.584	0.484	0.230	55.69
DEN	0.646	0.583	0.491	0.163	**81.14**
SIT	0.470	0.468	0.371	0.134	59.58
$\text{DEN}\|\|\text{TXT}_{ref}$	**0.654**	**0.632**	**0.531**	0.207	79.94
TXT_{web}	0.799	0.708	0.578	0.262	90.42

Table 1: Word similarity and relatedness evaluation

Model	TXT	DEN	SIT	VIS
TXT	1	0.60	0.45	0.30
DEN	0.60	1	0.45	0.35
SIT	0.45	0.35	1	0.26
VIS	0.30	0.35	0.26	1

Table 2: Model correlations

Table 2 shows correlations on cosine similarities on all word pairs from MEN, SemSim, VisSim and Compatibility between our word embeddings. This further corroborates the finding that different dimensions of grounding lead to complementary notions of similarity. In particular, correlation between visual and situational embeddings is relatively low, as compared to more fuzzy textual embeddings which correlate well with denotational embeddings. For a qualitative analysis, more examples are shown in Appendix A.

Qualitative Discussion Table 3 illustrates similarities learned from different grounding dimensions by means of some qualitative examples. Whereas denotational and visual embeddings rank semantically compatible words on top (e.g. *grass-grassy*), situational embeddings clearly focus more on topical similarity (*grass-clouds*). Given these examples, the finding that visual embeddings outperform denotational embeddings on the semantic compatibility task (see Table 1) seems rather contradictory. A preliminary error analysis suggests that the compatibility ratings that humans provide 'out of context' in a rating task differ

woman	txt$_{ref}$	lady, girl, man, chick
	den	lady, girl, women, blouse
	sit	girl, guy, man, lady
	vis	lady, girl, women, chick
sidewalk	txt$_{ref}$	pavement, ground, walkway, steps
	den	street, sidewlak, walkway, pavement
	sit	buildin, bldg, lamppost, street
	vis	pavement, street, walkway, concrete
grass	txt$_{ref}$	shrubs, dirt, bushes, sand
	den	grassy, patch, bounded, plains
	sit	clouds, church, trees, building
	vis	grassy, path, shrubs, bushes
couch	txt$_{ref}$	sofa, chair, bench, bed
	den	sofa, pillows, cushions, loveseat
	sit	sofa, leather, armchair, seater
	vis	sofa, pillow, pillows, love

Table 3: Top nearest neighbours for some example noun embeddings

to some extent from referential choices in our corpus. As an example, in the compatibility data set, the words *pigeon* and *mother* are rated as being equally similar to *animal*. However, in our corpus of referring expressions, *mother* is never used to refer to animal entities and our denotational embeddings predict them to be highly dissimilar, whereas visual embeddings are slightly more robust in this case.

More generally, textual embeddings learned from referring expressions captures a much more fuzzy and generic notion of similarity than denotational, visual or situational embeddings, e.g. *grass* is similar to *shrubs* and to *sand* in the textual space. This fuzziness has been found for word embeddings trained on large amounts of raw text as well (Faruqui et al., 2016).

4 Approximate Co-Reference Detection

Another important testbed for models of lexical meaning is their ability to capture semantic inference, with textual entailment as a well-known paradigm: here the task is to predict whether a textual hypothesis h can be inferred from a given premise p (Dagan et al., 2006). Young et al. (2014) have proposed a less strict variant of this called "approximate textual entailment". The main idea is that premise and hypothesis candidates can be automatically extracted from a corpus of captioned images. Given a set of captions known to describe the same image and an hypothesis, the task is to determine whether the hypothesis can describe the same image as the premise.

Inspired by this approach, we use the multimodal corpus of referring expressions to set up a new task for evaluating word embeddings, which consists of capturing approximate inferential relations between referring expressions. Thus, in our case, the hypothesis and the premise are expressions referring to objects, and the task is to determine whether they could (potentially) refer to the same object. Note that this is also similar to the notion of semantic compatibility proposed by Kruszewski and Baroni (2015), but extended to phrasal expressions. We can automatically extract positive and negative pairs from the data (see Section 2) by looking at pairs of expressions referring to objects in the same image and distinguishing **coreferential expressions** referring to the same entity (e.g. *grandma - old lady*), and **non-coreferential expressions** referring to different entities, e.g. *old lady - young lady*. In contrast to the majority of existing similarity and relatedness benchmarks which are centered around nouns, this task requires precise meaning representations for attribute-like words (e.g. *left-right*, *old-young*) which occur frequently in our data and which are frequently used to distinguish between objects occurring in the same situation. In particular, as the scenes in our data sets contain many objects of the same category (e.g. in the REFCOCO data), the distinction can often not be made by looking at the noun only, e.g. for classifying '*old* lady' - '*young* lady' as non-coreferential.

We call this task *approximate* coreference detection as the premise and hypothesis might describe complementary aspects of the same object such that the distinction cannot be made perfectly without the original perceptual context. For instance, in some cases, *lady in blue* and *young lady* might denote the same referent, in others not (see Figure 1). Thus, we note that the upper bound for automatic (or human) performance in this task is clearly not 100%. In future work, we plan to combine this with a reference resolution system that grounds the expressions in a given image.

Data and Set-up Given an image with several objects and a set of expressions referring to these, we compute the set of expression pairs P for that image. This set now divides into positive instances, i.e. expressions that both refer to the same object in the image, and negative instances, i.e. expressions that describe distinct entities in the scene. As this gives us a lot of data, we adopt a supervised learning approach for modeling the task of approximate co-reference detection. Thus, we use our embeddings to extract a range of similarity

measures between the expression pairs and feed these metrics as features into a classifier, trained to predict whether two phrases co-refer. This set-up is largely similar to Young et al. (2014)'s evaluation setting for approximate textual entailment.

Similarity Measures Given a pair P of expressions $r_i = w_{i_1} \dots w_{i_n}, r_j = w_{j_1} \dots w_{j_m}$, we extract pairwise cosine similarities between the embeddings $cos(w_{i_x}, w_{j_y})$, using average ($\sum_{(w_i,w_j)\in P} cos(w_i, w_j) \times \frac{1}{|P|}$), maximum ($\max_{(w_i,w_j)\in P} cos(w_i, w_j)$) and minimum distance ($\min_{(w_i,w_j)\in P} cos(w_i, w_j)$) as features for classification. Furthermore, we restrict the words in each expression such that they are disjunct sets excluding words that occur in both expressions, $w_i \neq w_j, \forall(w_i, w_j) \in P$. We extract the same average, maximum and minimum distance measures on these lexically disjunct expressions. Finally, we compose word embeddings for each expressions via addition ($r_i = w_{i_1} + \dots + w_{i_n}$) and add the cosine between the composed embeddings ($cos(r_i, r_j)$) to our list of features. Here, we compare textual, visual and denotational embeddings, as our situational embeddings only cover nouns.

Training From REFERIT, we extract 161K training and 18K test pairs, dividing into 66% non-coreferential and 34% coreferential expressions. We re-train our embeddings on the training portions of this data. We only consider non-coreferential expressions that refer to objects of the same type, according to their label annotated in the data set. From REFCOCO, we extract 300k pairs from the training set and 95k pairs from the test set, dividing into roughly 70% non-coreferential and 30% coreferential expressions. We randomly sample these pairs, the overall number of possible pairs in REFCOCO exceeds 2 million. We train a binary logistic regression classifier on each corpus, given the similarity measures extracted for each word embedding.

Results We report accuracies on co-referential expression detection in Table 4, on REFERIT and REFCOCO. Similarities derived from denotational embeddings clearly outperform the other classifiers on both data sets, including state-of-the-art textual embeddings learned on a much larger web corpus. On REFCOCO, only denotational embeddings lead to a clear improvement over the majority baseline. While the low performance of standard distributional embeddings is rather expected

	ReferIt	RefCoco
Majority	66.05	71.64
VIS	70.14	71.63
TXT_{ref}	68.49	71.57
DEN	**73.67**	**74.32**
TXT_{web}	69.16	71.89

Table 4: Accuracies for co-referential expression detection

top	txt_{ref}	upper, bototm, bottom, bottem
	den	upper, topmost, tippy, above
	vis	upper, above, of, corner
red	text	yellow, purple, maroon, blue
	den	maroon, redman, reddish, allmiddle
	vis	and, purple, yellow, pink
small	txt_{ref}	large, smaller, big, tiny
	den	smaller, smallest, little, littiest
	vis	directly, of, between, slightly

Table 5: Top nearest neighbours for some example adjectives embeddings

on this task (see previous findings on e.g. predicting antonyms (Nguyen et al., 2016)), the clear advange of denotational over visual embeddings is noteworthy. Whereas visual grounding is relatively effective for modeling compatibility between nouns (see Table 1), it does not seem to capture attribute meaning accurately as illustrated in Table 5. Here, the average of all visual objects referred to as e.g. *small* seems to be rather noisy and lead to high similarity with rather random words (*directly*) whereas denotational embeddings model accurate compatibility relations between e.g. *small-smaller*.

5 Conclusion

Whereas it is notoriously difficult to tailor or specialise distributional meaning representations inferred from text to particular aspects of semantic relatedness (Kiela et al., 2015a; Nguyen et al., 2016; Rimell et al., 2017), this work has shown that a multi-modal corpus of referring expressions can be used to derive a range of continuous meaning representations grounded in different aspects of context, capturing different notions of similarity. As compared to visual embeddings used in previous works, we found that denotational embeddings are particularly useful for detecting semantic relations. Other, recently proposed tasks related to modeling word association (Vulić et al., 2017), commonsense knowledge (Vedantam et al., 2015) or child-directed input (Lazaridou et al., 2016) provide interesting testbeds 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 *ACL (1)*, pages 238–247.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Anne Cocos and Chris Callison-Burch. 2017. The language of place: Semantic value from geospatial context. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 99–104. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

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*.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proc. of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

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*, pages 91–99. Association for Computational Linguistics.

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.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*.

Douwe Kiela, Felix Hill, and Stephen Clark. 2015a. Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048, Lisbon, Portugal. Association for Computational Linguistics.

Douwe Kiela, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015b. Exploiting image generality for lexical entailment detection. 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 119–124, Beijing, China. 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.

Germán Kruszewski and Marco Baroni. 2015. So similar and yet incompatible: Toward the automated identification of semantically compatible words. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 964–969, Denver, Colorado. Association for Computational Linguistics.

Angeliki Lazaridou, Grzegorz Chrupała, Raquel Fernández, and Marco Baroni. 2016. Multimodal semantic learning from child-directed input. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 387–392, San Diego, California. Association for Computational Linguistics.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. 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*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. 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.

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.

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)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 3111–3119, Lake Tahoe, Nevada, USA.

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.

Laura Rimell, Amandla Mabona, Luana Bulat, and Douwe Kiela. 2017. Learning to negate adjectives with bilinear models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 71–78. 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 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, page To appear.

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)*, pages 721–732, Baltimore, Maryland. 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.

Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2542–2550.

Ivan Vulić, Douwe Kiela, and Anna Korhonen. 2017. Evaluation by association: A systematic study of quantitative word association evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 163–175. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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