Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5865–5870
Marseille, 11–16 May 2020

Representing Verbs with Visual Argument Vectors

Irene Sucameli, Alessandro Lenci
CoLing Lab, University of Pisa
via S. Maria 36 - 56126 Pisa (Italy)
irene.sucameli@phd.unipi.it, alessandro.lenci@unipi.it

Abstract

Is it possible to use images to model verb semantic similarities? Starting from this core question, we developed two textual distributional semantic models and a visual one. We found it particularly interesting and challenging to investigate this Part of Speech since verbs are not often analysed in researches focused on multimodal distributional semantics. After the creation of the visual and textual distributional space, the three models were evaluated in relation to SimLex-999, a gold standard resource. Through this evaluation, we demonstrate that, using visual distributional models, it is possible to extract meaningful information and to effectively capture the semantic similarity between verbs.

Keywords: Distributional Semantic, Visual Models, verbs

1. Introduction

Imagine that someone offers you something to *drink*. Even if you do not know what kind of drink it might be, you could make some hypotheses. You could infer that the drink would be contained in a bottle or in a cup and that it would be either water, wine or beer. In the same way, you would consider unrealistic the idea of drinking an apple pie or a piece of chair. This type of educated guess is possible because you are able to derive the semantics of the verb *drink* from its linguistic and experiential context. For example, *water* often cooccurs with *drink* both in linguistic data, such as books, and in our multimodal experience.

Distributional Semantics is based on the assumption that words with similar meaning tend to be used within the same linguistic contexts. This hypothesis is implemented by Distributional Semantic Models (DSMs), which represent each lexical element through a distributional vector (Lenci, 2018). The similarity of two words is then calculated based on the position of their vectors within the distribution space. Purely linguistic DSMs may be extended to Multimodal DSMs (MDSMs), which combine the information coded within the linguistic vector with different type of data, the most common of which are visual information extracted from datasets of images (Lazaridou et al., 2015). These models tend to perform better than the ones based only on linguistic data, which show several limitations. In fact, even if linguistic models are able to capture complex linguistic properties, they do not register attributes that for a human being are absolutely intuitive. For example, it is more probable to extract from a text the information that a *lemon* is *sour* than that it is *yellow*, because it is unlikely that a writer would describe such obvious information as *lemons are yellow* (Baroni and Lenci, 2008; Andrews et al., 2009; Riordan and Jones, 2011). Similarly, if we would like to define an example of obvious information which is not encoded in a verb vector, we could imagine there is an higher probability to find sentences like ”John is punching the punching bag” instead of ”John is punching using his hands”. Thus, this information cannot be derived from traditional DSMs, while it can be represented with MDSMs.

In this paper, we propose a comparison of the performance obtained using DSMs based on textual information with the one obtained using visual DSMs that encode information extracted from images with the Bag of Visual Words (BoVW) technique (Bruni et al., 2011). This way we want to demonstrate that a visual distributional model is able to effectively capture the semantic similarity between concepts, performing in some cases even better than linguistic models.

Our analysis focuses in particular on verbs since they play a core role within the sentence. In fact, verbs more than nouns and adjectives are able to convey relevant information about events and actions described in sentences, and impose syntactic and semantics constraints on their arguments. Moreover, verbs have received little attention in MDSMs, which have mostly focused on nouns. The present research is based on the assumption that the multimodal representation of a verb can be derived from the visual representation of the verb's argument nouns and, more specifically, from those nouns which co-occur with the verb as subjects and objects.

The present paper is organized as follows. A short set of related works on multimodal Distributional Semantics are reviewed in Section 2. In Section 3, we describe the linguistic and visual resources used, as well as the methodology adopted for the creation of our DSMs. In Section 4, we discuss the analysis and the evaluation performed in order to determine whether the multimodal approach improved the model performance in capturing verb semantic similarity. Finally, Section 5 reports conclusions and ideas for further works.

2. Multimodal Distributional Semantics

Several works have been developed in years in multimodal Distributional Semantics. Here, we briefly mention the work of Bruni et al. (2014), one of the earliest on this subject. In this work, the textual and visual information are combined, producing a multimodal semantic model. Although their research is not the first attempt to combine textual and visual information, since Feng and Lapata already developed in 2010 a multimodal distributional semantic model using an approach which unified visual and textual information (Feng and Lapata, 2010), from the research of Bruni et al. emerges an interesting behavior of

MDSMs. In fact, the authors underline that while the models based on images are more oriented towards capturing the similarities between concrete nouns, focusing on properties such as colour or shape, textual models are more oriented towards recognizing abstract objects and their properties. According to Lazaridou et al. (2015) this limit of MDSMs may be due to the type of images which are extracted for abstract concepts. In their research, which is based on a multimodal version of Skip-gram (Mikolov et al., 2013b), the authors underline the necessity to extract more diversified images to represent abstract nouns such as *freedom* or *theory*. The reason for this is that, while for concrete words it is more common to find a picture that exactly represents that word, for abstract concepts this is more rare. As a consequence, the multimodal vectors of abstract word convey more complex information, since the visual data for an abstract concept can be extremely diversified one from another.

The difficulty to capture abstract properties and concepts using visual models is highlighted also in the research of Făgărăsan et al. (2015), who work on a method for the automatic prediction of the (visual) features of objects elicited by the subjects, and in the work by Hill and Korhonen (2014). In this latter research, authors focus their study on the development of a multimodal model to learn concrete as well as abstract nouns. These works reveal that visual models are able to extract different types of information with respect to textual DSMs. Köper and Schulte im Walde (2017) analyze the performance of a neural network model in predicting the compositionality of nominal and verbal multiword expressions, when visual information are included. From their results it emerges that, when combined with textual vectors that describe nouns, the visual features are able to predict better the concrete targets. Otherwise, when the textual verb vectors are combined with the features extracted from the images that visually describe the verb, the opposite case occurs. This demonstrates that the performance of MDSMs differs when applied to the study of nouns rather than to verbs.

Another interesting work is the one conducted by Shekhar et al. (2017a), which highlights an important limit of MDSMs. Although these models are able to recognize with a good level of accuracy the objects (represented linguistically by nouns) present in an image, they often have some difficulties in representing attributes (described by adjectives), actions (verbs), mode (adverbs) and spatial relations (prepositions). Adopting the FOIL methodology (Shekhar et al., 2017b), which consists in replacing a word in a generated caption with an incorrect element (the foil), Shekhar et al. demonstrate that MDSMs are often not able to completely identify all the elements present in an image. Therefore, according to the authors, it is compulsory to create models able to understand the information encoded by adjectives and prepositions.

Starting from these assumptions, the question we investigate in this paper is whether the information extracted through visual models is useful for defining the meaning of a verb.

3. A Visual Representation of Verbs

The aim of our study is to show that purely visual vectors can be used to effectively capture the semantic similarity between verbs. More specifically, in the present experiment textual verb vectors are syntax-based distributional representations that encode co-occurrences with its argument nouns, in particular subjects and objects.

In a similar way, visual verb vectors are built from the images that describe the subjects and objects of the verb. Starting from this assumption we i) selected the visual and the textual resources and ii) extracted the verb vectors from these resources.

3.1. The textual vectors

To build the textual vectors, we extracted the verbs from *SimLex-999* (Hill et al., 2015)[1] and their arguments from the annotated tensor of *Distributional Memory* (DM) (Baroni and Lenci, 2010)[2].

SimLex-999 is a resource which describes the similarity between pairs of words and is designed for the evaluation of DSMs. It includes 999 word pairs divided by POS (nouns, verbs and adjectives) and for two categories (concreteness and abstractness). Compared to other resources such as WordSim-353 (Finkelstein et al., 2002), SimLex-999 quantifies the similarity between pairs of words rather than their degree of association. This means that pairs related but not actually similar tend to have a lower score, compared to the one recorded in other datasets. For example, in SimLex-999 the pair *coast-shore* with a score of 9.00, while the pair *clothes-closet*, which are related but not similar, has a score of 1.96. In addition to this, Hill et al. remark on the difficulty DSMs have in capturing the similarity between verbs in SimLex. However, this feature is in line with the theory that verbs are relational concepts; thus, their meaning is more complex to grasp because it is closely linked to the one of the other words which co-occur within the sentence. Nevertheless, we decided to use this resource since it allows us to represent similarity between word pairs and to perform a detailed analysis of DSMs. From this dataset we randomly selected 100 target verbs (organized in 66 pairs), characterised by different degrees of concreteness. Then, for each target, we extracted its arguments from the DM tensor.

In DM, distributional information is organized in a set of weighted word-link-word tuples, formally represented as a third-order tensor. From the model *TypeDM*, we extracted the noun arguments of our target verbs, selecting the tuples marked with one of these links: *sbj-intr* (subject of a verb without a direct object), *sbj-tr* (subject of a verb with a direct object), *iobj* (indirect object), *obj* (direct object).

The extracted data were used to create two different DSMs: the first one (**M1**) consists of the 100 SimLex verbs as targets and their 20 noun arguments (10 subjects and 10 objects) with the highest value of Local Mutual Information as contexts, while the second one (**M2**) includes all their subjects and objects available in TypeDM. This means that

[1] Available at: `https://www.cl.cam.ac.uk/~fh295/simlex.html`

[2] Available at: `https://marcobaroni.org/dm/`

M1's nouns form a shorter list of M2's nouns, although both nouns arguments are marked with the same labels (*obj, iobj, sbj-tr, sbj-intr*). Then, to improve the quality of the semantic space the textual vectors were reduced to 100 latent dimensions with SVD. The reason why we chose to select 100 dimensions is that, as described by Bullinaria and Levy. (Bullinaria and Levy, 2012), the lower the number of dimensions considered the more efficient the computational distribution model should be. Nevertheless, different SVD dimensions and higher numbers of arguments extracted are been taking under consideration for future works.

3.2. The visual vectors

As visual resource we used *Imagenet*[3] (Deng et al., 2009), an ontology of images based on the same hierarchical structure used by WordNet (Miller, 2016). ImageNet is a rich dataset with diversified, high resolution images and a high level of image labels accuracy. We used an image corpus instead of a video corpus since our aim was to visually describe only the nominal occurrences of the verbs selected, while a video corpus may be more likely used to describe verbs and actions. To collect the images for our research, we selected the nominal occurrences from M1 and, from those 2,000 nouns, we identified 706 types. Then, we extracted the visual representations of the types and computed their BoVW using MMFeat (Kiela, 2016).

MMFeat[4] is a toolkit designed to simplify the extraction and the analysis of visual and audio resources for NLP tasks. With this toolkit, we downloaded randomly 5 images which visually described our target nouns, and computed their BoVW using the SIFT descriptor (Lowe, 1999). The SIFT descriptors are automatically calculated based on the most important features of the image and offer the advantage of remaining unchanged despite any alteration in light, position or point of view.

This way, the images representing the verb arguments used for the textual DSMs M1 and M2 were extracted. However, not all types of nouns are available in ImageNet. In fact, for some of them it was not possible to obtain a visual representation because they belong to the class of abstract nouns (such as *theory* and *experience*). Anyway, the problematic nature of finding visual representations from abstract nouns is attested in other researches (Hill and Korhonen, 2014; Anderson et al., 2017). Therefore, this result was quite expected. Instead, for the concrete arguments, it was possible to extract the corresponding images, identifying the descriptors and extracting the Bag of Visual Words. Then, the Visual Words that represented the same concept were aggregated and we calculated the centroid vector for each group of images. This way, we reduced the values of a Visual Word to a single value per image/concept.

At this point, as for the textual models, we organized the visual representation of verb's subjects and objects into a co-occurrence matrix. The visual matrix **MV** tables the verbs in rows, the images in columns and, as entries, the BOVW's centroids calculated per each image/concept. Thus, the visual vector associated with each verb is represented as the sequence of the centroids of its visual arguments. Finally, the visual matrix was also reduced to 100 latent dimension with SVD.

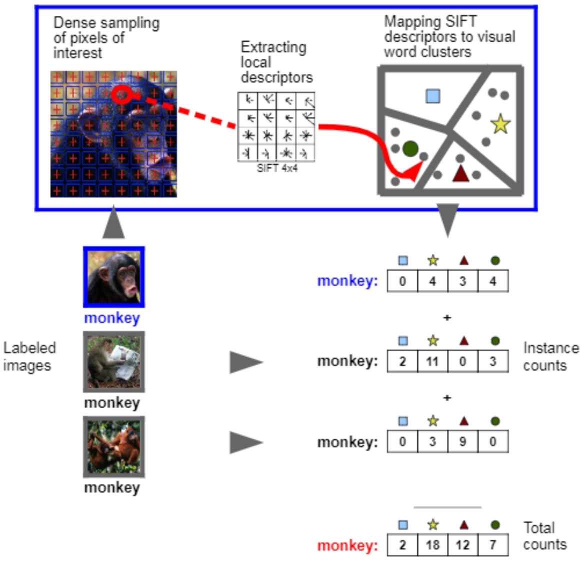

Figure 1: Example of how the image-based vector is constructed (Bruni et al., 2014)

4. Evaluation and Discussion

4.1. Evaluation with SimLex-999

In order to evaluate whether the visual distribution model was able to identify the similarity between verbs, we compared our models with the gold standard resource SimLex-999. The evaluation of our distributional models with SimLex-999 presents two difficulties. The first one is that, as highlighted also by Hill et al. (2015), distributional semantic models tend to perform worse when compared with this resource, because they are not usually able to identify the similarity between two words independently from the words' relatedness degree. The second difficulty depends on the type of PoS we decided to investigate. In fact, distributional semantic models perform poorly in recognizing the semantic similarity between verbs with respect to the others PoS.

The evaluation of M1, M2 and MV considers the gold standard resource's characteristic; so, since in SimLex-999 assigns low similarity score to pairs of antonyms, we did not include in the evaluation the antonym verb pairs. Our visual model MV obtained a good level of performance, with a Spearman's index of correlation ρ = **0.25** (Table 1). This result is highly positive if we consider that usually visual models tend to perform worse than the the textual ones, given the same number of co-occurrences (Bruni et al., 2014). Instead, our textual model M1, which had the same number of co-occurrences as MV, recorded a lower index of correlation ($\rho = 0.06$) than the visual distributional model realized.

Moreover, the results obtained by MV are competitive even when compared with external models. In fact, in the evaluation conducted by Hill et al. (2015) the standard Skip-

[3] Available at: http://www.image-net.org/index

[4] Available at: https://github.com/douwekiela/mmfeat

gram model (Mikolov et al., 2013a) obtained a correlation of $\rho = 0.27$ on the verb subset of SimLex-999 (Table 1).

	M1	*M2*	*MV*	*Skip-gram* (verbs)
SimLex-999	0.06	0.38	0.25	0.27

Table 1: Performances of our models and the Skip-gram model, as reported in Hill et al. (2015) for the verb subset of SimLex-999

Instead, it is clear that M2 performs better than MV ($\rho = 0.38$). This improvement is due mainly to the increase in the number of co-occurrences considered within the model. Indeed, as highlighted also by the authors of SimLex-999, models with more co-occurrences achieve better performances.

4.2. Abstract vs concrete

As we have already explained, the use of abstract words does affect the overall performance of visual distributional models, regardless of the POS considered (Hill and Korhonen, 2014; Făgărăsan et al., 2015; Anderson et al., 2017).
Therefore, we decided to make a further comparison between SimLex and the three models created, taking into consideration only concrete verbs. We evaluated the concreteness of a verb based on the study conducted by Power with respect to the categories of abstract verbs within the English language (Power, 2007); more specifically, Power illustrates two classes of abstract verbs and their characteristics. We then used Power's guidelines to define if a verb could be considered concrete or not. Should a verb falls in both categories (e.g. *have an idea* vs *have a car*) , we included it in the list of concrete verbs.
In comparison with the results obtained considering both abstract and concrete verbs (Figure 2), this second analysis shows that visual models can record a significant improvement when only concrete items are considered (Figure 3). On the other hand, the textual models M1 and M2 do not obtain a great improvement from this restriction. In fact, M1 presents just a little improvement (with a $\rho = 0.09$), M2 actually performs worse than before (ρ = 0.31) while MV records a index of correlation of ρ = 0.4 (while before it was 0.25).

This is an extremely interesting outcome from which it emerges that, augmenting the level of concreteness of the target data provided, the visual distributional model is able to perform better than before. More importantly, it emerges that using images is more useful in the definition of the semantic similarity between concrete verbs than using their corresponding linguistic counterparts. In fact, our concrete MV performed better than the textual models, describing successfully the similarity between verbs. Consequently, this implies that textual models achieve good performance scores only when both concrete and abstract verbs are included.
From the results obtained it is possible to conclude that, at least for the set of verbs considered, the concrete images have been proved to be more informative than the nouns denoting them.

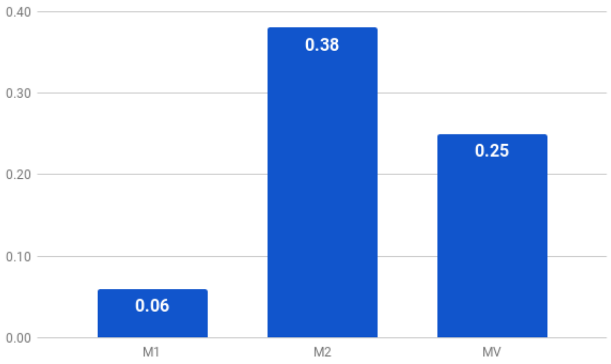

Figure 2: Evaluation of the three models compared to the gold standard when both concrete and abstracts verbs are considered (for a total of 100 verbs).

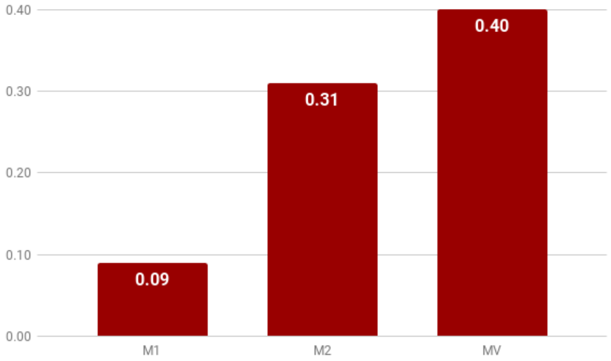

Figure 3: Evaluation of the three models compared to the gold standard when a subset of concrete verbs (45 verbs) is selected from SimLex-999.

5. Conclusion

In this paper we illustrated the development and the evaluation of a visual DSM applied to the study of the semantic similarity between verbs. The aim of our project was to demonstrate that using visual distributional models it is possible to effectively capture the semantic similarity between concepts, and that the information extracted through visual models is useful in capturing the meaning of a verb. To prove that, we created two textual distributional models, named M1 and M2, which included respectively the 20 top nominal co-occurrences (10 with the role of subject and 10 as object), and all the subjects and objects existing in DM for a set of target verbs selected from SimLex-999. We also built a visual model, MV, using the images extracted from ImageNet. Then, these three models have been compared in order to verify whether the visual model actually produced an improvement in defining the semantic similarity between verbs. From the results obtained, we saw that: i) our visual distributional model is actually able to capture the meaning of verbs and their semantic similarity, ii) performing even better than M1, which included the same number of nominal occurrences considered. However, if we compare two models with a different number of occurrences considered, the model with more occurrences is more able to identify the semantic similarities between verbs. The visual distributional model realised performed well if compared with SimLex-999, a gold standard resource whose use presents several critical issues. This data is particularly positive if we consider that i) visual models tend to have worse performances than the textual ones, and that ii) DSMs tend to perform poorly in recognizing the similarity between verbs respect with respect to nouns. Moreover, we showed that our visual model can obtain competitive results even when compared with external models, such as the Skip-gram model. Finally, we showed that the performance of the visual model can have a significant improvement, if we focus on concrete verbs only. In fact, visual models are heavily affected by the presence of abstract elements which are difficult to encode in an image. However, visual DSMs can describe the semantic similarity between concrete verbs better than textual models.

Having obtained positive results from the visual model presented in this paper, we intend to improve our approach, increasing the number of target verbs considered. With this regard, we are considering to use SimVerb-3500 (Gerz et al., 2016) since this is a larger verb-specific dataset. Moreover, we would like to increase the number of considered syntax-based co-occurrences as well as the number of vectors' dimension. Subsequently, this new and richer model could be used, in combination with a textual semantic model and representing the meaning of verbs with the concatenation of their textual and visual vectors, for the development of a novel MDSM.

6. Bibliographical References

Anderson, A., Kiela, D., Clark, S., and Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiental and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Baroni, M. and Lenci, A. (2008). Concept and properties in word spaces. *Italian Journal of Linguistics*, 20(1):55–88.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bruni, E., Tran, G., and Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, Edinburgh, UK.

Bruni, E., Tran, N., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(I):1–47.

Bullinaria, J. and Levy, J. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and svd. *Behavior Research Methods*, 44:890–907.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Făgărăsan, L., Vecchi, E. M., and Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57, London, UK.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, Los Angeles, CA.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *EMNLP (2016)*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Hill, F. and Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *EMNLP*, pages 255–265, Doha, Qatar.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(IV):665–695.

Kiela, D. (2016). Mmfeat: A toolkit for extracting multi-modal features. In *Proceedings of the 54th Annual Meeting of the Association for Computational LinguisticsâSystem Demonstrations*, pages 55–60, Berlin, Germany.

Köper, M. and Schulte im Walde, S. (2017). Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 200–206, Valencia, Spain.

Lazaridou, A., Pham, N., and Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 153–163. Denver, Colorado.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual review of Linguistics*, 4:151–171.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of International Conference of Computer Vision*, pages 1150–1157.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations*, pages 1–12, Scottsdale, Arizona, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS'13)*, pages 3111–3119.

Miller, G. (2016). Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Power, R. (2007). Abstract verbs. In *Proceedings of the 11th European Workshop on Natural Language Generation*, pages 93–96.

Riordan, B. and Jones, M. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):1–43.

Shekhar, R., Pezzelle, S., Herbelot, A., Nabi, M., Sangineto, E., and Bernardi, R. (2017a). Vision and language integration: Moving beyond objects. In *IWCS 2017*.

Shekhar, R., Pezzelle, S., Klimovich, Y., Herbelot, A., Nabi, M., Sangineto, E., and Bernardi, R. (2017b). Foil it! find one mismatch between image and language caption. In *ACL 2017*.