Journal of Artificial Intelligence Research 49 (2014) 1-47 Submitted 7/13; published 1/14

Multimodal Distributional Semantics

Elia Bruni ELIA.BRUNI@UNITN.IT
Center for Mind/Brain Sciences,
University of Trento, Italy

Nam Khanh Tran NTRAN@L3S.DE
L3S Research Center,
Hannover, Germany

Marco Baroni MARCO.BARONI@UNITN.IT
Center for Mind/Brain Sciences,
University of Trento, Italy
Department of Information Engineering and Computer Science,
University of Trento, Italy

Abstract

Distributional semantic models derive computational representations of word meaning from the patterns of co-occurrence of words in text. Such models have been a success story of computational linguistics, being able to provide reliable estimates of semantic relatedness for the many semantic tasks requiring them. However, distributional models extract meaning information exclusively from text, which is an extremely impoverished basis compared to the rich perceptual sources that ground human semantic knowledge. We address the lack of perceptual grounding of distributional models by exploiting computer vision techniques that automatically identify discrete "visual words" in images, so that the distributional representation of a word can be extended to also encompass its co-occurrence with the visual words of images it is associated with. We propose a flexible architecture to integrate text- and image-based distributional information, and we show in a set of empirical tests that our integrated model is superior to the purely text-based approach, and it provides somewhat complementary semantic information with respect to the latter.

1. Introduction

The **distributional hypothesis** states that words that occur in similar contexts are semantically similar. The claim has multiple theoretical roots in psychology, structuralist linguistics, lexicography and possibly even in the later writings of Wittgenstein (Firth, 1957; Harris, 1954; Miller & Charles, 1991; Wittgenstein, 1953). However, the distributional hypothesis has had a huge impact on computational linguistics in the last two decades mainly for empirical reasons, that is, because it suggests a simple and practical method to harvest word meaning representations on a large scale: Just record the contexts in which words occur in easy-to-assemble large collections of texts (corpora) and use their contextual profiles as surrogates of their meaning. Nearly all contemporary corpus-based approaches to semantics rely on contextual evidence in one way or another, but the most systematic and extensive application of distributional methods is found in what we call **distributional semantic models** (DSMs), also known in the literature as vector space or semantic space models of meaning (Landauer & Dumais, 1997; Sahlgren, 2006; Schütze, 1997; Turney & Pantel, 2010).

In DSMs, the meaning of a word is approximated with a vector that keeps track of the patterns of co-occurrence of the word in text corpora, so that the degree of semantic similarity or, more generally, relatedness (Budanitsky & Hirst, 2006) of two or more words can be precisely quantified in terms of geometric distance between the vectors representing them. For example, both *car* and *automobile* might occur with terms such as *street*, *gas* and *driver*, and thus their distributional vectors are likely to be very close, cuing the fact that these words are synonyms. Extended empirical evidence has shown that distributional semantics is very good at harvesting effective meaning representations on a large scale, confirming the validity of the distributional hypothesis (see some references in Section 2.1 below).

Still, for all its successes, distributional semantics suffers of the obvious limitation that it represents the meaning of a word entirely in terms of connections to other words. A long tradition of studies in cognitive science and philosophy has stressed how models where the meaning of symbols (e.g., words) are entirely accounted for in terms of other symbols (e.g., other words) without links to the outside world (e.g., via perception) are deeply problematic, an issue that is often referred to as the **symbol grounding problem** (Harnad, 1990). DSMs have also come under attack for their lack of grounding (Glenberg & Robertson, 2000).[1] Although the specific criticisms vented at them might not be entirely well-founded (Burgess, 2000), there can be little doubt that the limitation to textual contexts makes DSMs very dissimilar from humans, who, thanks to their senses, have access to rich sources of perceptual knowledge when learning the meaning of words – so much so that some cognitive scientists have argued that meaning is directly *embodied* in sensory-motor processing (see the work in de Vega, Glenberg, & Graesser, 2008, for different views on embodiment in cognitive science). Indeed, in the last decades a large amount of behavioural and neuroscientific evidence has been amassed indicating that our knowledge of words and concepts is inextricably linked with our perceptual and motor systems. For example, perceiving action-denoting verbs such as *kick* or *lick* involves the activation of areas of the brain controlling foot and tongue movements, respectively (Pulvermueller, 2005). Hansen, Olkkonen, Walter, and Gegenfurtner (2006) asked subjects to adjust the color of fruit images objects until they appeared achromatic. The objects were generally adjusted until their color was shifted away from the subjects' gray point in a direction opposite to the typical color of the fruit, e.g., bananas were shifted towards blue because subjects' overcorrected for their typical yellow color. Typical color also influences lexical access: For example, subjects are faster at naming a pumpkin in a picture in which it is presented in orange than in a grayscale representation, slowest if it is in another color (Therriault, Yaxley, & Zwaan, 2009). As a final example, Kaschak, Madden, Therriault, Yaxley, Aveyard, Blanchard, and Zwaan (2005) found that subjects are slower at processing a sentence describing an action if the sentence is presented concurrently to a visual stimulus depicting motion in the opposite

1. Harnard, in the original paper, is discussing formal symbols, such as those postulated in Fodor's "language of thought" (Fodor, 1975), rather than the words of a natural language. However, when the latter are represented in terms of connections to other words, as is the case in DSMs, the same grounding problem arises, and we follow the recent literature on the issue in referring to it as "*symbol* grounding", where our symbols are natural language words.

direction of that described (e.g., *The car approached you* is harder to process concurrently to the perception of motion away from you). See the review in Barsalou (2008) for a review of more evidence that conceptual and linguistic competence is strongly embodied.

One might argue that the concerns about DSMs not being grounded or embodied are exaggerated, because they overlook the fact that the patterns of linguistic co-occurrence exploited by DSMs reflect semantic knowledge we acquired through perception, so that linguistic and perceptual information are strongly correlated (Louwerse, 2011). Because dogs are more often brown than pink, we are more likely to talk about *brown dogs* than *pink dogs*. Consequently, a child can learn useful facts about the meaning of the concept denoted by *dog* both by direct perception and through linguistic input (this explains, among other things, why congenitally blind subjects can have an excellent knowledge of color terms; see, e.g., Connolly, Gleitman, & Thompson-Schill, 2007). One could then hypothesize that the meaning representations extracted from text corpora are indistinguishable from those derived from perception, making grounding redundant. However, there is by now a fairly extensive literature showing that this is not the case. Many studies (Andrews, Vigliocco, & Vinson, 2009; Baroni, Barbu, Murphy, & Poesio, 2010; Baroni & Lenci, 2008; Riordan & Jones, 2011) have underlined how text-derived DSMs capture encyclopedic, functional and discourse-related properties of word meanings, but tend to miss their concrete aspects. Intuitively, we might harvest from text the information that *bananas* are *tropical* and *eatable*, but not that they are *yellow* (because few authors will write down obvious statements such as "*bananas are yellow*"). On the other hand, the same studies show how, when humans are asked to describe concepts, the features they produce (equivalent in a sense to the contextual features exploited by DSMs) are preponderantly of a perceptual nature: *Bananas* are *yellow*, *tigers* have *stripes*, and so on.[2]

This discrepancy between DSMs and humans is not, *per se*, a proof that DSMs will face empirical difficulties as computational semantic models. However, if we are interested in the potential implications of DSMs as models of how humans acquire and use language –as is the case for many DSM developers (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996, and many others)– then their complete lack of grounding in perception is a serious blow to their psychological plausibility, and exposes them to all the criticism that classic ungrounded symbolic models have received. Even at the empirical level, it is reasonable to expect that DSMs enriched with perceptual information would outperform their purely textual counterparts: Useful computational semantic models must capture human semantic knowledge, and human semantic knowledge is strongly informed by perception.

If we accept that grounding DSMs into perception is a desirable avenue of research, we must ask where we can find a practical source of perceptual information to embed into DSMs. Several interesting recent experiments use features produced by human subjects in concept description tasks (so-called "semantic norms") as a surrogate of true perceptual features (Andrews et al., 2009; Johns & Jones, 2012; Silberer & Lapata, 2012; Steyvers, 2010). While this is a reasonable first step, and the integration methods proposed in these studies

2. To be perfectly fair, this tendency might in part be triggered by the fact that, when subjects are asked to describe concepts, they might be encouraged to focus on their perceptual aspects by the experimenters' instructions. For example McRae, Cree, Seidenberg, and McNorgan (2005) asked subjects to list first "physical properties, such as internal and external parts, and how [the object] looks."

are quite sophisticated, using subject-produced features is unsatisfactory both practically and theoretically (see however the work reported by Kievit-Kylar & Jones, 2011, for a crowdsourcing project that is addressing both kinds of concerns). Practically, using subject-generated properties limits experiments to those words that denote concepts described in semantic norms, and even large norms contain features for just a few hundred concepts. Theoretically, the features produced by subjects in concept description tasks are far removed from the sort of implicit perceptual features they are supposed to stand for. For example, since they are expressed in words, they are limited to what can be conveyed verbally. Moreover, subjects tend to produce only salient and distinctive properties. They do not state that dogs have a head, since that's hardly a distinctive feature for an animal!

In this article, we explore a more direct route to integrate perceptual information into DSMs. We exploit recent advances in computer vision (Grauman & Leibe, 2011) and the availability of documents that combine text and images to automatically extract visual features that are naturally co-occurring with words in multimodal corpora. These image-based features are then combined with standard text-based features to obtain perceptually-enhanced distributional vectors. In doing this, we rely on a natural extension of the distributional hypothesis, that encompasses not only similarity of linguistic context, but also similarity of visual context. Interestingly, Landauer and Dumais, in one of the classic papers that laid the groundwork for distributional semantics, already touched on the grounding issue and proposed, speculatively, a solution along the lines of the one we are implementing here: "[I]f one judiciously added numerous pictures of scenes with and without rabbits to the context columns in the [. . .] corpus matrix, and filled in a handful of appropriate cells in the *rabbit* and *hare* word rows, [a DSM] could easily learn that the words *rabbit* and *hare* go with pictures containing rabbits and not to ones without, and so forth." (Landauer & Dumais, 1997, p. 227).[3]

Although vision is just one source of perceptual data, it is a reasonable starting point, both for convenience (availability of suitable data to train the models) and because it is probably the dominating modality in determining word meaning. As just one piece of evidence for this claim, the widely used subject-generated semantic norms of McRae et al. (2005) contain 3,594 distinct perceptual features in total, and, of these, 3,099 (86%) are visual in nature!

Do the relatively low-level and noisy features that we extract from images in multimodal corpora contribute meaningful information to the distributional representation of word meaning? We report the results of a systematic comparison of the network of semantic relations entertained by a set of concrete nouns in the traditional text-based and novel image-based distributional spaces confirming that image-based features are, indeed, semantically meaningful. Moreover, as expected, they provide somewhat complementary information with respect to text-based features. Having thus found a practical and effective way to extract perceptual information, we must consider next how to combine text- and image-derived features to build a **multimodal distributional semantic model**. We propose a general parametrized architecture for multimodal fusion that, given appropriate sample data, automatically determines the optimal mixture of text- and image-based features to be used for the target semantic task. Finally, we evaluate our multimodal DSMs in

3. We thank Mike Jones for pointing out this interesting historical connection to us.

two separate semantic tasks, namely predicting the degree of semantic relatedness assigned to word pairs by humans, and categorizing nominal concepts into classes. We show that in both tasks multimodal DSMs consistently outperform purely textual models, confirming our supposition that, just like for humans, the performance of computational models of meaning improves once meaning is grounded in perception.

The article is structured as follows. Section 2 provides the relevant background from computational linguistics and image analysis, and discusses related work. We lay out a general architecture for multimodal fusion in distributional semantics in Section 3. The necessary implementation details are provided in Section 4. Section 5 presents the experiments in which we tested our approach. Section 6 concludes summarizing our current results as well as sketching what should come next.

2. Background and Related Work

In this section we first give a brief introduction to traditional distributional semantic models (i.e., those based solely on textual information). Then, we describe the image analysis techniques we adopt to extract and manipulate visual information. Next, we discuss earlier attempts to construct a multimodal distributional representation of meaning. Finally, we describe the most relevant strategies to combine information coming from text and images proposed inside the computer vision community.

2.1 Distributional Semantics

In the last few decades, a number of different distributional semantic models (DSMs) of word meaning have been proposed in computational linguistics, all relying on the assumption that word meaning can be learned directly from the linguistic environment.

Semantic space models are one of the most common types of DSM. They approximate the meaning of words with vectors that record their distributional history in a **corpus** (Turney & Pantel, 2010). A distributional semantic model is encoded in a matrix whose m rows are **semantic vectors** representing the meanings of a set of m **target words**. Each component of a semantic vector is a function of the occurrence counts of the corresponding target word in a certain context (see Lowe, 2001, for a formal treatment). Definitions of **context** range from simple ones (such as documents or the occurrence of another word inside a fixed window from the target word) to more linguistically sophisticated ones (such as the occurrence of certain words connected to the target by special syntactic relations) (Padó & Lapata, 2007; Sahlgren, 2006; Turney & Pantel, 2010). After the raw target-context counts are collected, they are transformed into **association scores** that typically discount the weights of components whose corresponding word-context pairs have a high probability of chance co-occurrence (Evert, 2005). The rank of the matrix containing the semantic vectors as rows can optionally be decreased by **dimensionality reduction**, that might provide beneficial smoothing by getting rid of noise components and/or allow more efficient storage and computation (Landauer & Dumais, 1997; Sahlgren, 2005; Schütze, 1997). Finally, the distributional semantic similarity of a pair of target words is estimated by a **similarity function** that takes their semantic vectors as input and returns a scalar similarity score as output.

There are many different semantic space models in the literature. Probably the best known is Latent Semantic Analysis (LSA, Landauer & Dumais, 1997), where a high-dimensional semantic space for words is derived by the use of co-occurrence information between words and the passages where they occur. Another well-known example is the Hyperspace Analog to Language model (HAL, Lund & Burgess, 1996), where each word is represented by a vector containing weighted co-occurrence values of that word with the other words in a fixed window. Other semantic space models rely on syntactic relations instead of windows (Grefenstette, 1994; Curran & Moens, 2002; Padó & Lapata, 2007). General overviews of semantic space models are provided by Clark (2013), Erk (2012), Manning and Schütze (1999), Sahlgren (2006) and Turney and Pantel (2010).

More recently, probabilistic topic models have been receiving increasing attention as an alternative implementation of DSMs (Blei, Ng, & Jordan, 2003; Griffiths et al., 2007). Probabilistic topic models also rely on co-occurrence information from large corpora to derive meaning but, differently from semantic space models, they are based on the assumption that words in a corpus exhibit some probabilistic structure connected to topics. Words are not represented as points in a high-dimensional space but as a probability distribution over a set of topics. Conversely, each topic can be defined as a probability distribution over different words. Probabilistic topic models tackle the problem of meaning representation by means of statistical inference: use the word corpus to infer the hidden topic structure.

Distributional semantic models, whether of the geometric or the probabilistic kind, ultimately are mainly used to provide a similarity score for arbitrary pairs of words, and that is how we will also employ them. Indeed, such models have shown to be very effective in modeling a wide range of semantic tasks including judgments of semantic relatedness and word categorization.

There are several data sets to assess how well a DSM captures human intuitions about semantic relatedness, such as the Rubenstein and Goodenough set (Rubenstein & Goodenough, 1965) and WordSim353 (Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman, & Ruppin, 2002). Usually they are constructed by asking subjects to rate a set of word pairs according to a similarity scale. Then, the average rating for each pair is taken as an estimate of the perceived relatedness between the words (e.g., *dollar-buck*: 9.22, *cord-smile*: 0.31). To measure how well a distributional model approximates human semantic intuitions, usually a correlation measure between the similarity scores generated by the model and the human ratings is computed. The highest correlation we are aware of on the WordSim353 set we will also employ below is of 0.80 and it was obtained by a model called Temporal Semantic Analysis, which captures patterns of word usage over time and where concepts are represented as time series over a corpus of temporally-ordered documents (Radinsky, Agichtein, Gabrilovich, & Markovitch, 2011). This temporal knowledge could be integrated with the perceptual knowledge we encode in our model. As a more direct comparison point, Agirre, Alfonseca, Hall, Kravalova, Pasça, and Soroa (2009) presented an extensive evaluation of distributional and WordNet-based semantic models on WordSim, both achieving a maximum correlation of 0.66 across various parameters.[4]

4. WordNet, available at `http://wordnet.princeton.edu/`, is a large computational lexicon of English where nouns, verbs, adjectives and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept.

Humans are very good at grouping together words (or the concepts they denote) into classes based on their semantic relatedness (Murphy, 2002), therefore a cognitive-aware representation of meaning must show its proficiency also in categorization (e.g., Poesio & Almuhareb, 2005; Baroni et al., 2010). Concept categorization is moreover useful for applications such as automated ontology construction and recognizing textual entailment. Unlike similarity ratings, categorization requires a discrete decision to group coordinates/cohyponyms into the same class and it is performed by applying standard clustering techniques to the model-generated vectors representing the words to be categorized. As an example, the Almuhareb-Poesio data set (Almuhareb & Poesio, 2005), that we also employ below, includes 402 concepts from WordNet, balanced in terms of frequency and degree of ambiguity. The distributional model of Rothenhäusler and Schütze (2009) exploits syntactic information to reach state-of-the-art performance on the Almuhareb-Poesio data set (maximum clustering purity across various parameter: 0.79). The window-based distributional approach of Baroni and Lenci (2010), more directly comparable to our text-based models, achieves 0.65 purity.

Other semantic tasks DSMs have been applied to include semantic priming, generation of salient properties of concepts and intuitions about the thematic fit of verb arguments (see, e.g., Baroni & Lenci, 2010; Baroni et al., 2010; McDonald & Brew, 2004; Padó & Lapata, 2007; Padó, Padó, & Erk, 2007). Distributional semantic vectors can be used in a wide range of applications that require a representation of word meaning, and in particular an objective measure of meaning relatedness, including document classification, clustering and retrieval, question answering, automatic thesaurus generation, word sense disambiguation, query expansion, textual advertising and some areas of machine translation (Dumais, 2003; Turney & Pantel, 2010).

2.2 Visual Words

Ideally, to build a multimodal DSM, we would like to extract visual information from images in a way that is similar to how we do it for text. Thanks to a well-known image analysis technique, namely bag-of-visual-words (**BoVW**), it is indeed possible to discretize the image content and produce visual units somehow comparable to words in text, known as **visual words** (Bosch, Zisserman, & Munoz, 2007; Csurka, Dance, Fan, Willamowski, & Bray, 2004; Nister & Stewenius, 2006; Sivic & Zisserman, 2003; Yang, Jiang, Hauptmann, & Ngo, 2007). Therefore, semantic vectors can be extracted from a corpus of images associated with the target (textual) words using a similar pipeline to what is commonly used to construct text-based vectors: Collect co-occurrence counts of target words and discrete image-based contexts (visual words), and approximate the semantic relatedness of two words by a similarity function over the visual words representing them.

The BoVW technique to extract visual word representations of documents was inspired by the traditional bag-of-words (BoW) method in Information Retrieval. BoW in turn is a dictionary-based method to represent a (textual) document as a “bag” (i.e., order is not considered), which contains words from the dictionary. BoVW extends this idea to visual documents (namely images), describing them as a collection of discrete regions, capturing their appearance and ignoring their spatial structure (the visual equivalent of ignoring word order in text). A bag-of-visual-word representation of an image is convenient from an image-

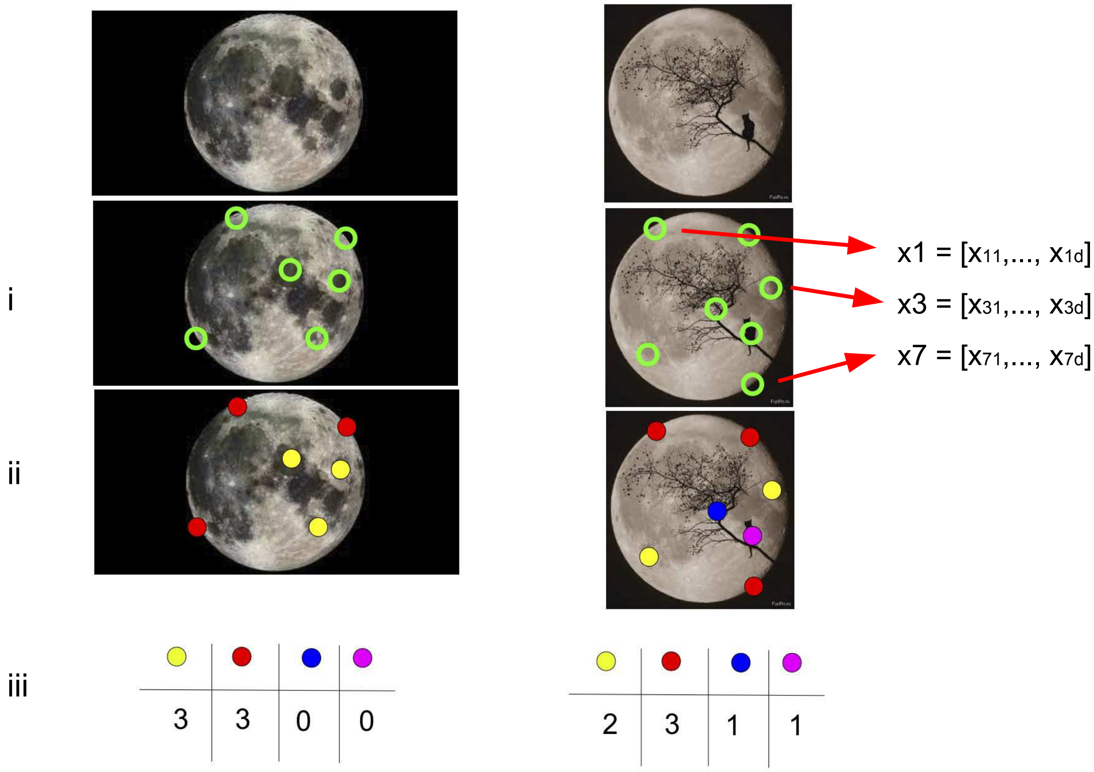

Figure 1: Representing images by BoVW: (i) Salient image patches or keypoints that contain rich local information are detected and represented as vectors of low-level features called descriptors; (ii) Descriptors are mapped to visual words on the basis of their distance from centers of clusters corresponding to the visual words (the preliminary clustering step is not shown in the figure); (iii) Images are finally represented as a bag-of-visual-words feature vector according to the distribution of visual words they contain. Images depicting the same things with rotations, occlusions, small differences in the low-level descriptors might still have a similar distribution of visual words, hence the same object can be traced very robustly across images while these conditions change.

analysis point of view because it translates a usually large set of high-dimensional local descriptors into a single sparse vector representation across images. Importantly, the size of the original set varies from image to image, while the bag-of-visual-word representation is of fixed dimensionality. Therefore, machine learning algorithms which by default expect fixed-dimensionality vectors as input (e.g., for supervised classification or unsupervised clustering) can be used to tackle typical image analysis tasks such as object recognition, image segmentation, video tracking, motion detection, etc.

More specifically, similarly to terms in a text document, an image has local interest points or **keypoints** defined as salient image patches that contain rich local information about the image. However "keypoint types" in images do not come off-the-shelf like word types in text documents. Local interest points have to be grouped into types (i.e., visual words) within and across images, so that an image can be represented by the number of occurrences of each type in it, analogously to BoW. The following pipeline is typically followed. From every image of a data set, keypoints are automatically detected (note that in most recent approaches a dense, pixelwise sampling of the keypoints is preferred to detecting the most salient ones only, and this is the solution that we also adopt, as explained in Section 4.2.2) and represented as vectors of low-level features called descriptors. Keypoint vectors are then grouped across images into a number of clusters based on their similarity in descriptor space. Each cluster is treated as a discrete visual word. With its keypoints mapped onto visual words, each image can then be represented as a BoVW feature vector recording how many times each visual word occurs in it. In this way, we move from representing the image by a varying number of high-dimensional keypoint descriptor vectors to a representation in terms of a single visual word count vector of fixed dimensionality across all images, with the advantages we discussed above. Visual word assignment and its use to represent the image content is exemplified in Figure 1, where two images with a similar content are described in terms of bag-of-visual-word vectors.

What kind of image content a visual word captures exactly depends on a number of factors, including the descriptors used to identify and represent keypoints, the clustering algorithm and the number of target visual words selected. In general, local interest points assigned to the same visual word tend to be patches with similar low-level appearance; but these local patterns need not be correlated with object-level parts present in the images (Grauman & Leibe, 2011).

2.3 Multimodal Distributional Semantics

The availability of large amounts of mixed media on the Web, on the one hand, and the discrete representation of images as visual words, on the other, has not escaped the attention of computational linguists interested in enriching distributional representations of word meaning with visual features.

Feng and Lapata (2010) propose the first multimodal distributional semantic model. Their generative probabilistic setting requires the extraction of textual and visual features from the same mixed-media corpus, because latent dimensions are here estimated through a probabilistic process which assumes that a document is generated by sampling both textual and visual words. Words are then represented by their distribution over a set of latent multimodal dimensions or "topics" (Griffiths et al., 2007) derived from the surface textual and visual features. Feng and Lapata experiment with a collection of documents downloaded from the BBC News website as corpus. They test their semantic representations on the free association norms of Nelson, McEvoy, and Schreiber (1998) and on a subset of 253 pairs from WordSim, obtaining gains in performance when visual information is taken into account (correlations with human judgments of 0.12 and 0.32 respectively), compared to the textual modality standalone (0.08 and 0.25 respectively), even if performance is still well below state-of-the-art for WordSim (see Section 2.1 above).

The main drawbacks of this approach are that the textual and visual data must be extracted from the same corpus, thus limiting the choice of the corpora to be used, and that the generative probabilistic approach, while elegant, does not allow much flexibility in how the two information channels are combined. Below, we re-implement the Feng and Lapata method (MixLDA) training it on the ESP-Game data set, the same source of labeled images we adopt for our model. This is possible because the data set contains both images and the textual labels describing them. More in general, we recapture Feng and Lapata's idea of a common latent semantic space in the latent multimodal mixing step of our pipeline (see Section 3.2.1 below).

Leong and Mihalcea (2011) also exploit textual and visual information to obtain a multimodal distributional semantic model. While Feng and Lapata merge the two sources of information by learning a joint semantic model, Leong and Mihalcea propose a strategy akin to what we will call Scoring Level fusion below: Come up with separate text- and image-based similarity estimates, and combine them to obtain the multimodal score. In particular, they use two combination methods: summing the scores and computing their harmonic mean. Differently from Feng and Lapata, Leong and Mihalcea extract visual information not from a corpus but from a manually coded resource, namely the ImageNet database (Deng, Dong, Socher, Li, & Fei-Fei, 2009), a large-scale ontology of images.[5] Using a handcoded annotated visual resource such as ImageNet faces the same sort of problems that using a manually developed lexical database such as WordNet faces with respect to textual information, that is, applications will be severely limited by ImageNet coverage (for example, ImageNet is currently restricted to nominal concepts), and the interest of the model as a computational simulation of word meaning acquisition from naturally occurring language and visual data is somewhat reduced (humans do not learn the meaning of "mountain" from a set of carefully annotated images of mountains with little else crowding or occluding the scene). In the evaluation, Leong and Mihalcea experiment with small subsets of WordSim, obtaining some improvements, although not at the same level we report (the highest reported correlation is 0.59 on just 56 word pairs). Furthermore they use the same data set to tune and test their models.

In Bruni, Tran, and Baroni (2011) we propose instead to directly concatenate the text- and image-based vectors to produce a single multimodal vector to represent words, as in what we call Feature Level fusion below. The text-based distributional vector representing a word, taken there from a state-of-the-art distributional semantic model (Baroni & Lenci, 2010), is concatenated with a vector representing the same word with visual features, extracted from all the images in the ESP-Game collection we also use here. We obtain promising performance on WordSim and other test sets, although appreciably lower than the results we report here (we obtain a maximum correlation of 0.52 when text- and image-based features are used together; compare to Table 2 below).

Attempts to use multimodal models derived from text and images to perform more specific semantic tasks have also been reported. Bergsma and Goebel (2011) use textual and image-based cues to model selectional preferences of verbs (which nouns are likely arguments of verbs). Their experiment shows that in several cases visual information is more useful than text in this task. For example, by looking in textual corpora for words such as *carillon*, *migas* or *mamey*, not much useful information is obtained to guess which of the three is a plausible argument for the verb *to eat*. On the other hand, they also show

5. `http://image-net.org/`

that, by exploiting Google image search functionality,[6] enough images for these words are found that a vision-based model of edible things can classify them correctly.

Finally, we evaluate our multimodal models in the task of discovering the color of concrete objects, showing that the relation between words denoting concrete things and their typical color is better captured when visual information is also taken into account (Bruni, Boleda, Baroni, & Tran, 2012). Moreover, we show that multimodality helps in distinguishing literal and nonliteral uses of color terms.

2.4 Multimodal Fusion

When textual information is used for image analysis, this is mostly done with different aims than ours: Text is used to improve image-related tasks, and typically there is an attempt to model the relation between specific images and specific words or textual passages (e.g., Barnard, Duygulu, Forsyth, de Freitas, Blei, & Jordan, 2003; Berg, Berg, & Shih, 2010; Farhadi, Hejrati, Sadeghi, Young, Rashtchian, Hockenmaier, & Forsyth, 2010; Griffin, Wahab, & Newell, 2013; Kulkarni, Premraj, Dhar, Li, Choi, Berg, & Berg, 2011). In contrast, (i) we want to use image-derived features to improve the representation of word meaning and (ii) we are interested in capturing the meaning of word *types* on the basis of sets of images connected to a word, and not to model specific word-image relations.

Despite these differences, some of the challenges addressed in the image analysis literature that deals with exploiting textual cues are similar to the ones we face. In particular, the problem of merging, or “fusing”, textual and visual cues into a common representational space is exactly the same we have to face when we construct a multimodal semantic space.

Traditionally, the image analysis community distinguishes between two classes of fusion schemes, namely early fusion and late fusion. The former fuses modalities in feature space, the latter fuses modalities in semantic similarity space, analogously to what we will call Feature Level and Scoring Level fusion, respectively. For example, Escalante, Hérnadez, Sucar, and Montes (2008) propose an image retrieval system for multimodal documents. Both early and late fusion strategies for the combination of the image and the textual channels are considered. Early fusion settings include a weighted linear combination of the two channels and a global strategy where different retrieval systems are used contemporarily on the entire, joint data set. Late fusion strategies include a per-modality strategy, where documents are retrieved by using only one or the other channel and a hierarchical setting where first text, image and their combination are used independently to query the database and then results are aggregated with four weighted combinations. Vreeswijk, Huurnink, and Smeulders (2011) train a visual concept classifier for abstract subject categories such as biology and history by using a late fusion approach where image and text information are combined at the output level, that is, first obtaining classification scores from the image- and text-based models separately and then joining them. Similarly to our multimodal mixing step, Pham, Maillot, Lim, and Chevallet (2007) and Caicedo, Ben-Abdallah, González, and Nasraoui (2012) propose an early fusion in which the two inputs are mapped onto the same latent space using dimensionality reduction techniques (e.g., Singular Value Decomposition). The multimodal representation obtained in this way is then directly used to retrieve image documents.

6. `http://images.google.com/`

3. A Framework for Multimodal Distributional Semantics

In this section, a general and flexible architecture for multimodal semantics is presented. The architecture makes use of distributional semantic models based on textual and visual information to build a multimodal representation of meaning. To merge the two sources, it uses a parameter-based pipeline which is able to capture previously proposed combination strategies, with the advantage of having all of them explored within a single system.

3.1 Input of the Multimodal Architecture

To construct a multimodal representation of meaning, a semantic model for each single modality has to be implemented. Independently of the actual parameters that are chosen for its creation (that, from our point of view, can be in a black box), there are some requirements that each model has to satisfy in order to guarantee a good functioning of the framework. In the first place, each modality must provide a separate representation, to leave room for the various fusion strategies afterwards. Then, each modality must encode the semantic information pertaining to each word of interest into a fixed-size vectorial representation. Moreover, we assume that both text- and image-based vectors are normalized and arranged in matrices where words are rows and co-occurring elements are columns.

In what follows, we assume that we harvested a matrix of text-based semantic vectors, and one of image-based semantic vectors for the same set of target words, representing, respectively, verbal and visual information about the words. In Section 4 below we give the details of how in our specific implementation we construct these matrices.

3.2 Multimodal Fusion

The pipeline is based on two main steps:

(1) Latent Multimodal Mixing: The text and vision matrices are concatenated, obtaining a single matrix whose row vectors are projected onto a single, common space to make them interact.

(2) Multimodal Similarity Estimation: Information in the text- and image-based matrices is combined in two ways to obtain similarity estimates for pairs of target words: at the Feature Level and at the Scoring Level.

Figure 2 describes the infrastructure we propose for fusion. First, we introduce a mixing phase to promote the interaction between modalities that we call Latent Multimodal Mixing. While this step is part of what other approaches would consider Feature Level fusion (see below), we keep it separated as it might benefit the Scoring Level fusion as well.

Once the mixing is performed, we proceed to integrate the textual and visual features. As reviewed in Section 2.4 above, in the literature fusion is performed at two main levels, the Feature Level and the Scoring Level. In the first case features are first combined and considered as a single input for operations, in the second case a task is performed separately with different sets of features and the separate results are then combined. Each approach has its own advantages and limitations and this is why both of them are incorporated into the multimodal infrastructure and together constitute what we call Multimodal Similarity

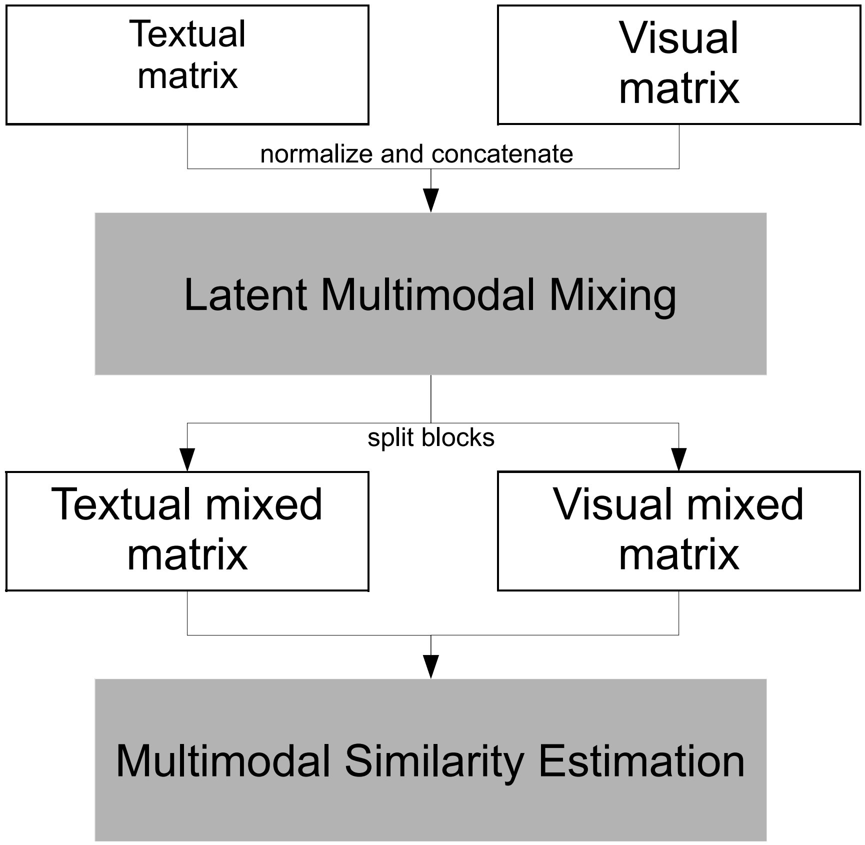

Figure 2: Multimodal fusion for combining textual and visual information in a semantic model.

Estimation. A Feature Level approach requires only one learning step (i.e., determining the parameters of the feature vector combination) and offers a richer vector-based representation of the combined information, that can also be used for other purposes (e.g., image and text features could be used together to train a classifier). Benefits of a Scoring Level approach include the possibility to have different representations (in principle, not even vectorial) and different similarity scores for different modalities and the ease of increasing (or decreasing) the number of different modalities used in the representation.

3.2.1 Latent Multimodal Mixing

This is a preparatory step in which the textual and the visual components are projected onto a common representation of lower dimensionality to discover correlated latent factors. The result is that new connections are made in each source matrix taking into account information and connections present in the other matrix, originating from patterns of co-variance that overlap. Importantly, we assume that mixing is done via a dimensionality reduction technique that has the following characteristics: a parameter k that determines the dimensionality of the reduced space and the fact that when k equals the rank of the original matrix the reduced matrix is identical or can be considered a good approximation of the original one. The commonly used Singular Value Decomposition reduction method that we adopt here for the mixing step satisfies these constraints.

As a toy example of why mixing might be beneficial, consider the concepts *pizza* and *coin*, that we could use as features in our text-based semantic vectors (i.e., record the co-occurrences of target words with these concepts as part of the vector dimensions). While these words are not likely to occur in similar contexts in text, they are obviously visually similar. So, the original text features *pizza* and *coin* might not be highly correlated. However, after mixing in multimodal space, they might both be associated with (have high weights on) the same reduced space component, if they both have similar distributions to visual features that cue roundness. Consequently, two textual features that were originally uncorrelated might be drawn closer to each other by multimodal mixing, if the corresponding concepts are visually similar, resulting in mixed textual features that are, in a sense, visually enriched, and *vice versa* for mixed visual features (interestingly, psychologists have shown that, under certain conditions, words such as *pizza* and *coin*, that are not strongly associated but perceptually similar, can prime each other; e.g., Pecher, Zeelenberg, & Raaijmakers, 1998).

Note that the matrices obtained by splitting the reduced-rank matrix back into the original textual and visual blocks have the same number of feature columns as the original textual and visual blocks, but the values in them have been smoothed by dimensionality reduction (we explain the details of how this is achieved in our specific implementation in the next paragraph). These matrices are then used to calculate a similarity score for a word pair by (re-)merging information at the feature and scoring levels.

Mixing with SVD In our implementation, we perform mixing across text- and image-based features by applying the Singular Value Decomposition (**SVD**)[7] to the matrix obtained by concatenating the two feature types row-wise (so that each row of the concatenated matrix describes a target word in textual and visual space). SVD is a widely used technique to find the best approximation of the original data points in a space of lower underlying dimensionality whose basis vectors ("principal components" or "latent dimensions") are selected to capture as much of the variance in the original space as possible (Manning, Raghavan, & Schütze, 2008, Ch. 18). By performing SVD on the concatenated textual and visual matrices, we project the two types of information into the same space, where they are described as linear combinations of principal components. Following the description by Pham et al. (2007), the SVD of a matrix M of rank r is a factorization of the form

$$M = U\Sigma V^t$$

where

$$\begin{cases} U : \text{matrix of eigenvectors derived from } MM^t \\ \Sigma : r \times r \text{ diagonal matrix of singular values } \sigma \\ \sigma : \text{square roots of the eigenvalues of } MM^t \\ V^t : \text{matrix of eigenvectors derived from } M^tM \end{cases}$$

7. Computed with SVDLIBC: `http://tedlab.mit.edu/~dr/SVDLIBC/`

In our context, the matrix M is given by normalizing two feature matrices separately and then concatenating. By selecting the k largest values from matrix Σ and keeping the corresponding columns in matrices U and V, the reduced matrix M_k is given by

$$M_k = U_k \Sigma_k V_k^t$$

where $k < r$ is the dimensionality of the latent space. While M_k keeps the same number of columns/dimensions as M, its rank is now k. k is a free parameter that we tune on the development sets. Note that when k equals the rank of the original matrix, then trivially $M_k = M$. Thus we can consider not performing any SVD reduction as a special case of SVD, which helps when searching for the optimal parameters.

Note also that, if M has n columns, then V_k^t is a $k \times n$ matrix, so that M_k has the same number of columns of M. If the first j columns of M contain textual features, and columns from $j+1$ to n contain visual features, the same will hold for M_k, although in the latter the values of the features will have been affected by global SVD smoothing. Thus, in the current implementation of the pipeline in Figure 2, block splitting is attained simply by dividing M_k into a textual mixed matrix containing its first j columns, and a visual mixed matrix containing the remaining columns.

3.2.2 Multimodal Similarity Estimation

Similarity Function Following the distributional hypothesis, DSMs describe a word in terms of the contexts in which it occurs. Therefore, to measure the similarity of two words DSMs need a function capable of determining the similarity of two such descriptions (i.e., of two semantic vectors). In the literature, there are many different similarity functions used to compare two semantic vectors, including cosine similarity, Euclidean distance, L_1 norm, Jaccard's coefficient, Jensen-Shannon divergence, Lin's similarity. For an extensive evaluation of different similarity measures, see the work by Weeds (2003).

Here we focus on **cosine** similarity since it has been shown to be a very effective measure on many semantic benchmarks (Bullinaria & Levy, 2007; Padó & Lapata, 2007). Also, given that our system is based on geometric principles, the cosine, together with Euclidean distance, is the most principled choice to measure similarity. For example, some of the measures listed above, having been developed from probabilistic considerations, will only be applicable to vectors that encode well-formed probability distributions, which is typically not the case (for example, after multimodal mixing, our vectors might contain negative values).

The cosine of two semantic vectors **a** and **b** is their dot product divided by the product of their lengths:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{i=n} a_i \times b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

The cosine ranges from 0 (orthogonal vectors) to $|1|$ (parallel vectors pointing in the same or opposite directions have cosine values of 1 and -1, respectively).

Feature Level Fusion In Feature Level fusion (**FL**), we use the linear weighted fusion method to combine text- and image-based feature vectors of words into a single representation and then we use the latter to estimate the similarity of pairs. The linear weighted combination function is defined as

$$F = \alpha \times F_t \oplus (1 - \alpha) \times F_v$$

where $\oplus$ is the vector-concatenate operator.

Scoring Level Fusion In Scoring Level fusion (**SL**), text- and image-based matrices are used to estimate similarity of pairs independently. The scores are then combined to obtain the final estimate by using a linear weighted scoring function:

$$S = \beta \times S_t + (1 - \beta) \times S_v$$

General Form and Special Cases Given fixed and normalized text- and image-based matrices, our multimodal approach is parametrized by k (dimensionality of latent space), FL vs. SL, α (weight of text component in FL similarity estimation) and β (weight of text component in SL).

Note that when k=r, with r the rank of the original combined matrix, Latent Multimodal Mixing returns the original combined matrix (no actual mixing). Picking SL with β=1 or β=0 corresponds to using the textual or visual matrix only, respectively. We thus derive as special cases the models in which only text (k=r, SL, β=1) or only images (k=r, SL, β=0) are used (called **Text** and **Image** models in the Results section below). The simple approach of Bruni et al. (2011), in which the two matrices are concatenated without mixing, is the parametrization k=r, FL, α=0.5 (called **NaiveFL** model, below). The summing approach of Leong and Mihalcea (2011) corresponds to k=r, SL, β=0.5 (**NaiveSL**, below). Picking $k<r$, SL, $\beta=1$ amounts to performing latent multimodal mixing, but then using textual features only; and the reverse with mixed image features only for $\beta=0$ (**Text**$_{mixed}$ and **Image**$_{mixed}$, respectively). Reducing these and other models to the same parametrized approach means that, given a development set for a specific task that requires similarity measurements, we can discover in a data-driven way which of the various models is best for the task at hand (for example, for a certain task we might discover that we are better off using text only, for another mixed text features, for yet another both text and image features, and so on).

Formally, given the set $k_1, ..., k_n \in \mathbb{R}$ of n dimensionalities of the latent space (with k_n equal to the original dimensionality, and arbitrary steps between the chosen values), the sets $\alpha_1, ..., \alpha_m \in \mathbb{R}$ of m potential weights of the text block in FL (with $\alpha_1 = 0$ and $\alpha_m = 1$) and $\beta_1, ..., \beta_l \in \mathbb{R}$ of l weights of the text block in SL (with $\beta_1 = 0$ and $\beta_l = 1$), we can calculate the number of possible configurations to explore by $tot_c = n(m + l)$. Unless n, m and l are very large (i.e., we consider very small intervals between the values to be tested), it is completely feasible to perform a full search for the best parameters for a certain task without approximate optimization methods. In our experiments, $n = 9$, $m = l = 11$, and consequently $tot_c = 198$.

4. Implementation Details

Both our implementation of the multimodal framework and of the visual feature extraction procedure are publicly available and open source.[8] Moreover the visual feature extraction procedure is presented by Bruni, Bordignon, Liska, Uijlings, and Sergienya (2013).

4.1 Construction of the Text-Based Semantic Matrix

As reviewed in Section 2.1 above, a text-based distributional model is encoded in a matrix whose rows are "semantic vectors" representing the meaning of a set of target words. Important parameters of the model are the choice of **target** and **contextual elements**, the **source corpora** used to extract co-occurrence information, the **context** delimiting the scope of co-occurrence, and the function to transform raw counts into statistical **association scores** downplaying the impact of very frequent elements.

Source Corpora We collect co-occurrence counts from the concatenation of two corpora, ukWaC and Wackypedia (size: 1.9B and 820M running words, or tokens, respectively). ukWaC is a collection of Web pages based on a linguistically-controlled crawl of the `.uk` domain conducted in the mid 2000s. Wackypedia was built from a mid-2009 dump of the English Wikipedia. Both corpora have been automatically annotated with lemma (dictionary form) and part-of-speech (POS) category information using the TreeTagger,[9] they are freely and publicly available,[10] and they are widely used in linguistic research.

Target and Context Elements Since our source corpora are annotated with lemma and part-of-speech information, we take both into account when extracting target and context words (e.g., the string *sang* is treated as an instance of the verb lemma *sing*). We collect semantic vectors for a set of 30K target words (lemmas), namely the top 20K most frequent nouns, 5K most frequent adjectives and 5K most frequent verbs in the combined corpora. The same 30K lemmas are also employed as contextual elements (consequently, our text-based semantic models are encoded in a 30K×30K matrix). Note that when we combine the text matrices with the image-based ones, we preserve only those rows (target words) for which we also have an image-based vector, trimming the matrix to size 20,525×30K.

Context We define context in terms of words that co-occur within a window of fixed width, in the tradition of the popular HAL model (Lund & Burgess, 1996). Window-based models are attractive for their simplicity and the fact that they do not require resource-intensive advanced linguistic annotation. They have moreover been reported to be at the state of the art in various semantic tasks (Rapp, 2003; Sahlgren, 2008), and in Bruni, Uijlings, Baroni, and Sebe (2012) we show that the window-based methods we use here outperform both a document-as-context model and a sophisticated syntax- and lexical-pattern-based model on the MEN and WordSim test sets introduced in Section 5.2 below (see also the post-hoc analysis using the document-based model discussed at the end of Section 5.2.2 below). We consider two variants, **Window2** and **Window20** (we chose these particular variants arbitrarily, as representatives of narrow and wide windows, respectively).

8. See `https://github.com/s2m/FUSE/` and `https://github.com/vsem/`, respectively.
9. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
10. `http://wacky.sslmit.unibo.it/`

Window2 records sentence-internal co-occurrence with the nearest 2 content words to the left and right of each target word (function words such as articles and prepositions are ignored). Window20 considers a larger window of 20 content words to the left and right of the target. A narrower window is expected to capture a narrower kind of semantic similarity, such as the one that exists between terms that are closely taxonomically related, for example coordinate concepts (*dog* and *cat*) or pairs of superordinate and subordinate concepts (*animal* and *dog*). The rationale behind this expectation is that terms will share many narrow-window collocates only if they are *very* similar, both semantically and syntactically. On the other hand, a broader window will capture a broader kind of "topical" similarity, such as one would expect of words that tend to occur in the same paragraphs (for example, *war* and *oil*, that are rather distant concepts in a taxonomic sense, but might easily occur in the same discourse). See the work by Sahlgren (2006) for further discussion of the effects of context width on distributional semantic models.

Association Score We transform raw co-occurrence counts into nonnegative Local Mutual Information (**LMI**) association scores. LMI scores are obtained by multiplying raw counts by Pointwise Mutual Information, and in the nonnegative case they are a close approximation to Log-Likelihood Ratio scores, that are one of the most widely used weighting schemes in computational linguistics (Evert, 2005). The nonnegative LMI of target element t and context element c is defined as:

$$LMI(t,c) = \max\left(\mathrm{Count}(t,c) \times \log\frac{\mathrm{P}(t,c)}{\mathrm{P}(t)\mathrm{P}(c)}, 0\right)$$

It is worth observing that, in an extensive study of how parameters affect the quality of semantic vectors, Bullinaria and Levy (2007) and Bullinaria and Levy (2012) found that a model similar to our Window2 (co-occurrence statistics from ukWaC, narrow window, lemmatized content word collocates, nonnegative pointwise mutual information instead of LMI) performs at or near the top in a variety of semantic tasks. Thus, we have independent grounds to claim that we are using a state-of-the-art text-based model.

4.2 Construction of the Image-Based Semantic Matrix

Given that image-based semantic vectors are a novelty with respect to text-based ones, in the next subsections we dedicate more space to how we constructed them, including full details about the source corpus we utilize as input of our pipeline (Section 4.2.1), the particular image analysis technique we choose to extract visual collocates and how we finally arrange them into semantic vectors that constitute the visual block of our distributional semantic matrix (Section 4.2.2).

4.2.1 Image Source Corpus

We adopt as our source corpus the ESP-Game data set[11] that contains 100K images, labeled through the famous "game with a purpose" developed by Louis von Ahn, in which two

11. http://www.cs.cmu.edu/~biglou/resources/

Figure 3: Samples of images and their tags from the ESP-Game data set

people partnered online must independently and rapidly agree on an appropriate word to label random selected images. Once a word is entered by both partners in a certain number of game rounds, that word is added as a tag for that image, and it becomes a taboo term for next rounds of the game involving the same image, to encourage players to produce more terms describing the image (Von Ahn, 2006). The tags of images in the data set form a vocabulary of 20,515 distinct word types. Images have 14 tags on average (4.56 standard deviation), while a word is a tag for 70 images on average (737.71 standard deviation).

To have the words in the same format as in our text-based models, the tags are lemmatized and POS-tagged. To annotate the words with their parts of speech, we could not run a POS-tagger, since here words are out of context (i.e., each tag appears alphabetically within the small list of words labeling the same image and not within the ordinary sentence required by a POS-tagger). Thus we used a heuristic method, which assigned to the words in the ESP-Game vocabulary their most frequent tag in our textual corpora.

The ESP-Game corpus is an interesting data set from our point of view since, on the one hand, it is rather large and we know that the tags it contains are related to the images. On the other hand, it is not the product of experts labelling representative images, but of a noisy annotation process of often poor-quality or uninteresting images (e.g., logos) randomly downloaded from the Web. Thus, analogously to the characteristics of a textual corpus, our algorithms must be able to exploit large-scale statistical information, while being robust to noise. While cleaner and more illustrative examples of each concept are available in carefully constructed databases such as ImageNet (see Section 2.3), noisy tag annotations are available on a massive scale on sites such as Flickr[12] and Facebook,[13] so if we want to eventually exploit such data it is important that our methods can work on noisy input. A further advantage of ESP-Game with respect to ImageNet is that its images are associated not only with concrete noun categories but also with adjectives, verbs and nouns related to events (e.g., *vacation*, *party*, *travel*, etc). From a more practical point of view, "clean" data sets such as ImageNet are still relatively small, making experimentation with standard benchmarks difficult. In concrete, looking at the benchmarks we experiment with, as of mid 2013, ImageNet covers only just about half the pairs in the WordSim353 test set, and less than 40% of the Almuhareb-Poesio words. While in the future we want to explore to what extent higher-quality data sources can improve image-based models, this will require larger databases, or benchmarks relying on a very restricted vocabulary.

The image samples in Figure 6 exemplify different kinds of noise that characterize the ESP-Game data set. Both on top and bottom left and top right there are images where the scene is cluttered or partially occluded. The top center image is hardly a good representative of accompanying words such as *building*, *tower(s)* or *square*. Similarly, the center bottom image is only partially a good illustration of a coin, and certainly not a very good example of a man! Finally, the bottom right image is useless from a visual feature extraction perspective.

4.2.2 Image-Based Semantic Vector Construction

We collect co-occurrence counts of target words and image-based contexts by adopting the BoVW pipeline that, as we already explained in 2.2, is particularly convenient in order to discretize visual information into "visual collocates". We are adopting what is currently considered a standard implementation of BoVW. In the future, we could explore more cutting-edge ways to build image-based semantic vectors, such as local linear encoding (Wang, Yang, Yu, Lv, Huang, & Gong, 2010) or Fisher encoding (Perronnin, Sanchez, & Mensink, 2010). Chatfield, Lempitsky, Vedaldi, and Zisserman (2011) present a systematic evaluation of several recent methods.

Our current implementation is composed of the following steps: (i) Extraction of the **local descriptors**, that is, vectors of low-level features that encode geometric or other information about the area around each keypoint, i.e., pixel of interest (here, SIFT descriptors); (ii) **Constructing a vector representation of an image** by assigning the local descriptors to clusters corresponding to visual words, and recording their distribution across these clusters in the vector (this presupposes a preliminary step in which a clustering algorithm has been applied to the whole image collection or a sample, to determine the visual word vocabulary) (iii) Including some spatial information into the representation with **spatial binning**; (iv) Summing visual word occurrences across the list of images associated with a word label to obtain the **co-occurrence counts** associated with each word label and transforming these counts into association scores, analogously to what is done in text analysis. The process (without spatial binning) is schematically illustrated in Figure 4, for a hypothetical example in which there are three images in the collection labeled with the word *monkey*. More details follow.

12. `http://www.flickr.com`
13. `http://www.facebook.com`

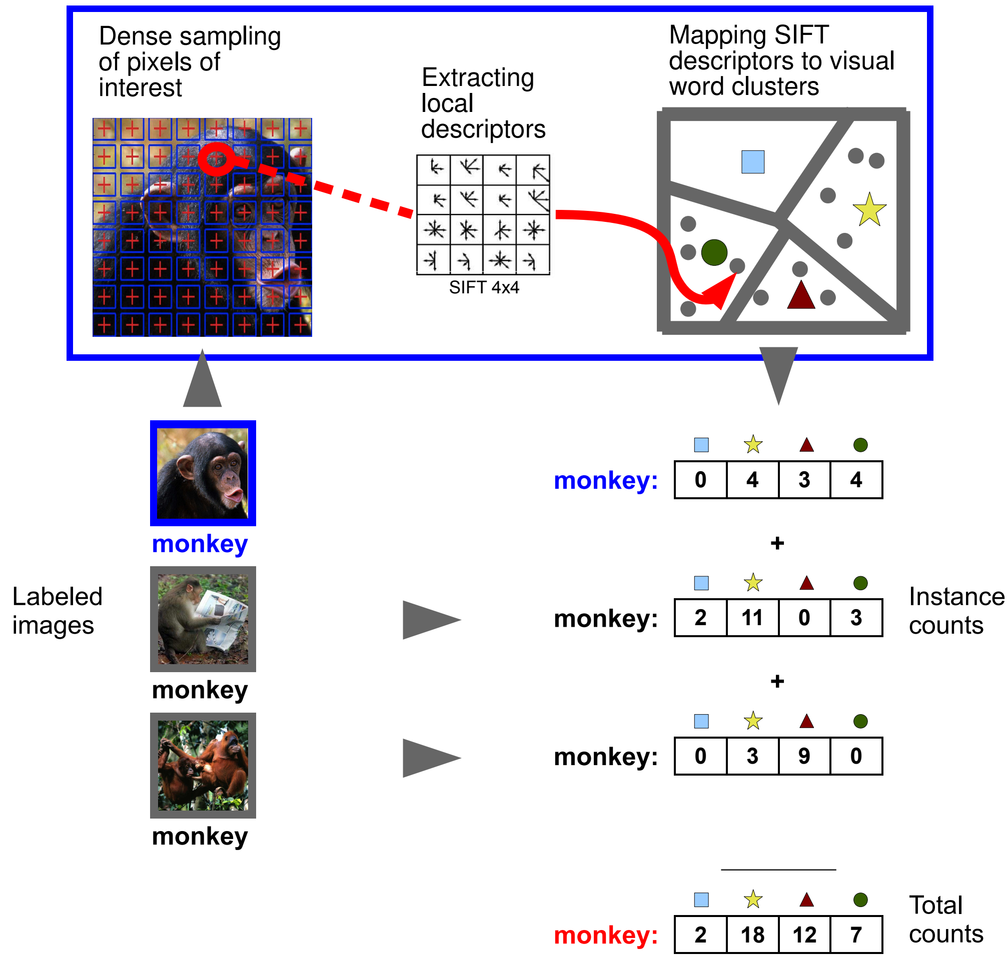

Figure 4: The procedure to build an image-based semantic vector for a target word. First, a bag-of-visual-word representation for each image labeled with the target word is computed (in this case, three images are labeled with the target word *monkey*). Then, the visual word occurrences across instance counts are summed to obtain the co-occurrence counts associated with the target word.

Local Descriptors To construct the local descriptors of pixels of interest we use Scale-Invariant Feature Transform (SIFT) (Lowe, 1999, 2004). We chose SIFT for its invariance to image scale, orientation, noise, distortion and partial invariance to illumination changes. A SIFT vector is formed by measuring the local image gradients in the region around each location and orientation of the feature at multiple scales. In particular, the contents of 4×4 sampling subregions are explored around each keypoint. For each of the resulting 16 samples, the magnitude of the gradients at 8 orientations are calculated, which would already result in a SIFT feature vector of 128 components. However, we extract color SIFT descriptors in HSV (Hue, Saturation and Value) space (Bosch, Zisserman, & Munoz, 2008). We use HSV because it encodes color information in a similar way to how humans do. We compute SIFT descriptors for each HSV component. This gives 3×128 dimensions per descriptor, 128 per channel. Color channels are then averaged to obtain the final 128-dimensional descriptors. We experimented also with different color scales, such as LUV, LAB and RGB, obtaining significantly worse performance compared to HSV on our development set introduced in 5.2.1, therefore we do not conduct further experiments with them. Van de Sande, Gevers, and Snoek (2010) present a systematic evaluation of color features.

Instead of searching for interesting keypoints with a salient patch detection algorithm, we use a more computationally intensive but also more thorough dense keypoint sampling approach, with patches of fixed size and localized on a regular grid covering the whole image and repeated over multiple scales. SIFT descriptors are computed on a regular grid every five pixels, at four scales (10, 15, 20, 25 pixel radii) and zeroing the low contrast descriptors. For their extraction we use the vl_phow command included in the VLFeat toolbox (Vedaldi & Fulkerson, 2010). This implementation has been shown to be very close to Lowe's original but it is much faster for dense feature extraction. Nowak, Jurie, and Triggs (2006) report a systematic evaluation of different patch sampling strategies.

Importantly, SIFT feature vectors are extracted from a large corpus of representative images to populate a feature space, which subsequently is quantized into a discrete number of visual words by clustering. Once this step is performed, every SIFT vector (local descriptor) from the original or new images can be translated into a visual word by determining which cluster it is nearest to in the quantized space.

Visual Vocabulary To map SIFT descriptors to visual words, we first cluster all local descriptors extracted from all images in a training image corpus in their 3×128-dimensional space using the k-means clustering algorithm, and encode each descriptor by the index of the cluster (visual word) to which it belongs. k-means is the most common way of constructing visual vocabularies (Grauman & Leibe, 2011). Given a set $\mathbf{x_1}, ..., \mathbf{x}_n \in R^D$ of n training descriptors, k-means aims to partition the n descriptors into k sets ($k \leq n$) so as to minimize the cumulative approximation error $\sum_{\mathbf{i=1}}^{n} ||\mathbf{x}_i - \mu_{q_\mathbf{i}}||^\mathbf{2}$, with K centroids $\mu_\mathbf{1}, ..., \mu_K \in R^D$ and data-to-means assignments $\mathbf{q_1}, ..., \mathbf{q}_N \in \{1, ..., K\}$. We use an approximated version of k-means called Lloyd's algorithm (1982) as implemented in the VLFeat toolbox.

To construct our visual vocabulary we extracted SIFT descriptors from all the 100K images of the ESP-Game data set. To tune the parameter k we used the MEN development set (see Section 5.2.1). By varying k between 500 and 5000 in steps of 500, we found the optimal k being 5000. It is most likely that the performance has not peaked even at 5000 visual words and enhancements could be attained by adopting larger visual vocabularies via more efficient implementations of the BoVW pipeline, as for example by Chatfield et al. (2011).

Image Representation Given a set of descriptors $\mathbf{x_1}, ..., \mathbf{x}_n$ sampled from an image, let q_i be the assignment of each descriptor $\mathbf{x}_i$ to its corresponding visual word. The bag-of-visual-words representation of an image is a nonnegative vector $v \in R^k$ such that $v_k = |\{i : q_i = k\}|$, with q ranging from 1 to the number of visual words in the vocabulary (in our case, 5000). This representation is a vector of visual words obtained via hard quantization (i.e., assignment of each local descriptor vector to the single nearest codeword).

Spatial Binning A consolidated way of introducing weak *geometry* in BoVW is the use of spatial histograms (Grauman & Darrell, 2005; Lazebnik, Schmid, & Ponce, 2006). The main idea is to divide the image in several (spatial) regions and to perform the entire visual word extraction and counting pipeline for each region and then concatenate the vectors. In our experiments the spatial regions are obtained by dividing the image in 4×4, for a total of 16 regions. Therefore, crossing the values for k with the spatial region, we increase the feature dimensions 16 times, for a total of 80,000 components in our vectors.

Co-occurrence Counts and Weighting Once the BoVW representations are built, each target (textual) word is associated to the list of images which are labeled with it; the visual word occurrences across the list of images is summed to obtain the co-occurrence counts associated with the target (textual) word. In total, 20,515 target words (those that constitute ESP-Game tags) have an image-based semantic vector associated.

Also in the image-based semantic matrix, like in the text-based one, raw counts are transformed into nonnegative LMI. The difference is that here LMI is computed between a target element t that is a textual word and a context element c that is a visual word instead.

Note that, just like in the standard textual approach, we are accumulating visual words from all images that contain a word without taking into account the fact that words might denote concepts with multiple appearances, can be polysemous or even hide homonyms (our *bank* vector will include visual words extracted from river as well as building pictures). An interesting direction for further research would be to cluster the images associated to a word in order to distinguish the "visual senses" of the word, e.g., along the lines of what was done for textual models by Reisinger and Mooney (2010).

4.3 Multimodal Fusion Tuning

We performed two separate parameter optimizations, one specifically for the semantic relatedness task (using MEN development, see Section 5.2.1) and the other specifically for the clustering task (using Battig, see Section 5.3.1). We determined the best model by performing an exhaustive search across SVD k (from 2^4 to 2^{12} in powers of 2), FL and SL with α varying from 0 to 1 (inclusive) in steps of 0.1 and similarly for β. In total, 198 models were explored and the one with the highest performance on the development data was chosen. Note that tuning was performed separately for the Window2 and Window20 models.

4.4 MixLDA

To reimplement Feng and Lapata's approach (discussed in Section 2.3) in a comparable setting to ours, we treat the ESP-Game data set as a mixed-media corpus where each image together with the associated tags constitutes a document. For each image, we extract the image-based features with the procedure described above in 4.2.2 and use the words labeling that image to obtain the text-based features. These features are then stored in a term-by-document matrix, in which each image is treated as a document and a term can be either a textual tag or a visual word extracted from that image. We obtain a matrix of size 90K$\times$100K, with 10K textual words (the word list resulting from the intersection of all the words used in our experimental data sets), 80K visual words and 100K documents (images).

The Latent Dirichlet Allocation (MixLDA) model is trained on this matrix and tuned on the MEN development set by varying the number of topics K_t.[14] The optimal value we find is $K_t = 128$. Under MixLDA, each target word in an evaluation set is represented by the vector giving its distribution over the 128 latent topics.

5. Experiments

We test our semantic representation in three different tasks, that is, evaluating the distribution of different kinds of semantic relations among a word's neighbours (5.1), modeling word relatedness judgments (5.2) and clustering words into superordinate concepts (5.3). Together, these tasks should give us a clear idea of the general quality of our models and of the relative contribution of visual information to meaning representation.

5.1 Differentiation Between Semantic Relations

To acquire a qualitative insight into how well our text- and image-based models are capturing word meaning, we test them on BLESS (Baroni-Lenci Evaluation of Semantic Similarity), a benchmark recently introduced by Baroni and Lenci (2011) to analyze specific aspects of lexico-semantic knowledge. Rather than focusing on a point estimate of quality of a model on a specific semantic task, BLESS allows us to assess the overall pattern of semantic relations that the model tends to capture. We run the BLESS evaluation before combining the textual and the visual channels together as a sanity check on the semantic meaningfulness of the image-based vectors, looking for potential complementary information with respect to text which can further motivate fusion. Note that since we are not combining the textual and visual sources, there are no tuning parameters to report.

5.1.1 Benchmark and Method

BLESS contains a set of 200 **pivot** words denoting concrete concepts (we use 184 pivots, since for the remaining 16 we do not have a sufficiently large set of related words covered by our models). For each of the pivots, the data set contains a number of related words, or **relata**, instantiating the following 8 common **semantic relations** with the pivots: COORD: the relatum is a noun that is a co-hyponym (coordinate) of the pivot (*alligator-lizard*); HYPER: the relatum is a noun that is a hypernym (superordinate) of the pivot (*alligator-reptile*); MERO: the relatum is a noun referring to a meronym, that is, a part or material of the pivot (*alligator-teeth*); ATTRI: the relatum is an adjective expressing an attribute of the pivot (*alligator-ferocious*); EVENT: the relatum is a verb referring to an action or event involving the concept (*alligator-swim*); RAN.N, RAN.J and RAN.V, finally, are control cases where the pivot is matched to a set of random nouns (*alligator-trombone*), adjectives (*alligator-electronic*) and verbs (*alligator-conclude*), respectively.

For each pivot, BLESS contains a set of relata of each category (ranging from 7 hypernyms to 33 random nouns per pivot on average). In this way, BLESS can highlight the broader semantic properties of a model independently of its more specific preferences. For example, both a model that assigns a high score to *alligator-ferocious* and a model that assigns a high score to *alligator-green* will be correctly treated as models that have picked

14. LDA was computed with Gensim: `http://radimrehurek.com/gensim/`

a relevant attribute of *alligators*. At the same time, the comparison of the specific relata selected by the models allows a more granular qualitative analysis of their differences.

Following the guidelines of Baroni and Lenci (2011), we analyze a semantic model as follows. We compute the cosine between the model vectors representing each of the 184 pivots and each of its relata, picking the relatum with the highest cosine for each of the 8 relations (the nearest hypernym, the nearest random noun, etc.). We then transform the 8 similarity scores collected in this way for each pivot onto standardized z scores (to get rid of pivot-specific effects), and produce a boxplot summarizing the distribution of scores per relation across the 184 pivots (for example, the leftmost box in the first panel of Figure 5 reports the distribution of 184 standardized cosines of nearest coordinate relata with the respective pivots). Besides analyzing the distributions qualitatively, we also discuss significant differences between the cosines of different relation types that were obtained via Tukey's Honestly Significance tests, thus correcting for multiple pairwise comparisons (Abdi & Williams, 2010).

5.1.2 Results

In Fig. 5, we report BLESS nearest relata distributions for the purely textual model Window20 (the Window2 distribution shows an even stronger skew in favour of coordinate neighbours) and the purely visual model we call Image in the next sections. The patterns produced by the text-based model (left panel) illustrate how a sensible word meaning profile should look like: coordinates are the most similar terms (an *alligator* is maximally similar to a *crocodile*), followed by superordinates (*reptile*) and parts (*teeth*). Semantically related adjectives (attri: *ferocious*) and verbs (event: *swim*) are less close to the pivots, but still more so than any random item.

The right panel shows the distribution of relata in the image-based semantic vectors. The overall pattern is quite similar to the one observed with the text-based vectors: there is a clear preference for coordinates, followed by hypernyms and parts, then attributes and events, with all random relata further away from the pivots than the semantically meaningful categories. For both models, coordinates are significantly closer to the relata than hypernyms and meronyms, that are significantly closer than attributes and events, that are in turn significantly closer than any random category. Although the difference between hypernyms and parts is not significant with either representation, intriguingly the image-based vectors show a slight preference for the more imageable parts (*teeth*) than the more abstract hypernyms (*reptile*). The only difference of statistical import is the one between events and attributes, where the text-based model shows a significant preference for events, whereas the two categories are statistically indistinguishable in the image-based model (as we will see shortly, the relative preference of the latter for attributes is probably due to its tendency to pick perceptual adjectives denoting color and size).

Looking more closely at the specific relata picked by the text- and image-based models, the most striking differences pertain, again, to attributes. The text- and image-based models picked the same attribute for a pivot in just 20% of the cases (compare to 40% overlap across all non-random relation types). Table 1 reports the attributes picked by the text- vs. image-based models for 20 random cases where the two mismatch.

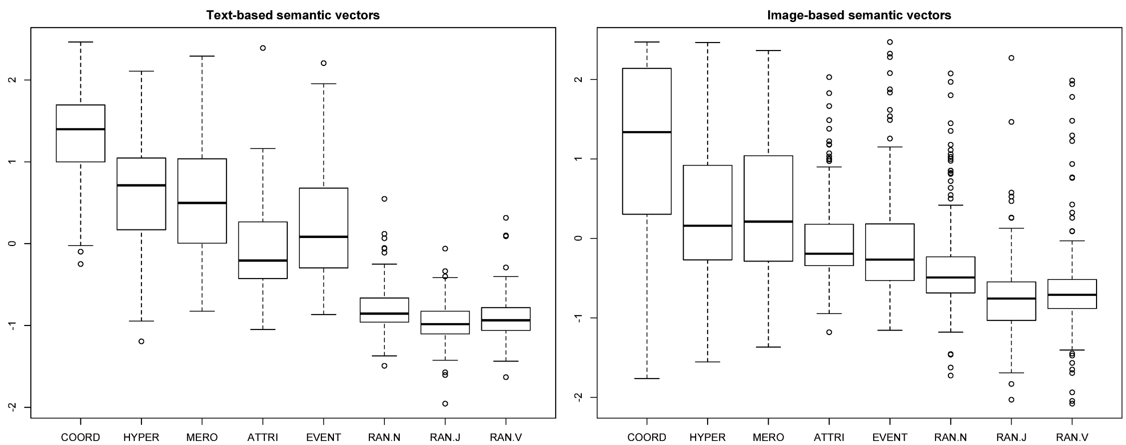

Figure 5: Distribution of z-normalized cosines of words instantiating various relations across BLESS pivots. Text-based vectors from the Window20 model.

pivot	*text*	*image*	*pivot*	*text*	*image*
cabbage	leafy	white	helicopter	heavy	old
carrot	fresh	orange	onion	fresh	white
cherry	ripe	red	oven	electric	new
deer	wild	brown	plum	juicy	red
dishwasher	electric	white	sofa	comfortable	old
elephant	wild	white	sparrow	wild	little
glider	heavy	white	stove	electric	hot
gorilla	wild	black	tanker	heavy	grey
hat	white	old	toaster	electric	new
hatchet	sharp	short	trout	fresh	old

Table 1: Attributes preferred by text- (Window20) vs. image-based models.

It is immediately clear from the table that, despite the fact that the pivots are nouns denoting concrete concepts, the text-based model almost never picks adjectives denoting salient perceptual properties (and in particular visual properties: just *white* for *hat* and *leafy* for *cabbage*). The text-based model focuses instead on encyclopedic properties such as *fresh, ripe, wild, electric* and *comfortable.* This is in line with earlier analyses of the "ungrounded" semantics provided by text-based models (Andrews et al., 2009; Baroni et al., 2010; Baroni & Lenci, 2008; Riordan & Jones, 2011), and differs greatly from the trend found in the image-based model. In 12/20 cases, the closest attribute for the latter model is a color. In the remaining cases, we have size (*short*, *little*), one instance of *hot* and, surprisingly, four of *old.*

To conclude, the analysis we presented confirms, on the one hand, our hypothesis that image-based distributional vectors contain sufficient information to capture a network of sensible word meaning relations. On the other, there are intriguing differences in the relations picked by the text- and image-based models, pointing to their complementarity.

5.2 Word Relatedness

As is standard in the distributional semantics literature (Budanitsky & Hirst, 2006; Sahlgren, 2006), we assess the performance of our models on the task of predicting the degree of semantic relatedness between two words as rated by human judges. We test the models on the WS and MEN benchmarks.

5.2.1 Benchmarks and Method

WS, that is, WordSim353[15] (see also Section 2.1) is a widely used benchmark constructed by asking 13 subjects to rate a set of 353 word pairs on an 11-point meaning similarity scale and averaging their ratings (e.g., *dollar/buck* gets a very high average rating, *professor/cucumber* a very low one). Our target words cover 252 WS pairs (thus, the correlations reported below are not directly comparable to those reported in other studies that used WS). However, our text-based models have much higher WS coverage (96%). When evaluated on the larger WS set they cover, Window2 and Window20 achieve 0.64 and 0.68 correlations, respectively. We are thus comparing the multimodal approach with purely textual models that are at the state of the art for WS (see results reported in Section 2.1 above).

The second benchmark we use, **MEN** (for Marco, Elia and Nam, the resource creators) was developed by us, specifically for the purpose of testing multimodal models. We created a large data set that, while comparable to WS and other benchmarks commonly used by the computational semantics community, contains only words that appear as image labels in the ESP-Game and MIRFLICKR-1M[16] collections, thus ensuring full coverage to researchers that train visual models from these resources. MEN consists of 3,000 word pairs with $[0, 1]$-normalized semantic relatedness ratings provided by Amazon Mechanical Turk workers (via the CrowdFlower[17] interface). For example, *beach/sand* has a MEN score of 0.96, *bakery/zebra* received a 0 score.

15. `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`
16. `http://press.liacs.nl/mirflickr/`
17. `http://crowdflower.com/`

Compared to WS, MEN is sufficiently large to allow us to separate development and test data, avoiding issues of overfitting. We use indeed 2,000 MEN pairs (development set) for model tuning and 1,000 pairs for evaluation (test set). Importantly, the development set has been used to find the best configuration once for both the MEN test set and WS. Thus, the WS evaluation illustrates how well the parameters learned on training data from a specific data set generalize when applied to the same semantic task but on a different data set.

Models are evaluated as follows. For each pair in a data set, we compute the cosine of the model vectors representing the words in the pair, and then calculate the Spearman correlation of these cosines with the (pooled) human ratings of the same pairs, the idea being that the higher the correlation the better the model can simulate the relatedness scores.

MEN Construction An earlier version of MEN has been used for the first time by the authors by Bruni et al. (2012) but since the current article is the first major publication in which we focus specifically on it, and we have recently improved the benchmark by extending the ratings, we provide here further details on how it was constructed.

The word pairs that constitute MEN were randomly selected from words that occur at least 700 times in the concatenated ukWaC and Wackypedia text corpora and at least 50 times as tags in the ESP-Game and MIRFLICKR-1M tagged image collections. In order to avoid picking only pairs that were weakly related, as would happen if we were to sample random word pairs from a list, we ranked all possible pairs by their cosines according to our text-based model Window20. To gather the 3000 word pairs needed for the construction of MEN, we subsequently picked the first 1000 word pairs, another 1000 was sampled from pairs placed between 1001 and 3000 in the cosine-ranked list and the last block of 1000 pairs from the remaining items.

To acquire human semantic relatedness judgments, we decided to ask for comparative judgments on two pair exemplars at a time rather than absolute scores for single pairs, as was done by the creators of WS. This should constitute a more natural way to evaluate the target pairs, since human judgments are comparative in nature. When a person evaluates a given target, she does not do so in a vacuum, but in relation with a certain context. Moreover, binary choices were preferred because they make the construction of “right” and “wrong” control items straightforward (see Footnote 18). Operationally, each word pair was randomly matched with a comparison pair coming from the same set of 3000 items and rated by a single Turker as either more or less related than the comparison item. The validity of this approach is confirmed by the high annotation accuracy we observe in the control set,[18] and by the high correlation of the MEN scores with ratings collected on a Likert scale we report below.

18. The control items are correct annotations created prior to running the job on Amazon Mechanical Turk, which act as hidden tests that are randomly shown to Turkers as they complete the job. In this way, we can calculate the quality of a contributor’s performance and reject their annotations if the accuracy drops below a certain percentage (we set a required minimum precision equal to 70%, but we obtained almost 100% average accuracy overall). Control items are also of great help to train quickly new workers to perform the required task. To create our control items we harvested two equally-sized sets of word pairs from WS, one containing only pairs with a high relatedness score, one containing only pairs with a low relatedness score. Each control item was then obtained by juxtaposing a high score pair with a low score pair and by treating the pair with the higher score as the one that should be selected by the

In the instructions, annotators were warned that sometimes both candidate pairs could contain words related in meaning and in such cases we asked them to pick the pair with the more strongly related words (e.g., both *wheels-car* and *dog-race* are somewhat related pairs, but the first one should be preferred as every car has wheels but not every dog is involved in a race). In other cases, annotators could find that neither pair contains closely related words, and in such cases they were instructed to pick the pair that contained slightly more related words (e.g., neither *auction-car* nor *cup-asphalt* are closely related words, but the first pair should be picked because fancy vintage cars are sold at auctions). We requested participants to be native speakers and only accepted those connecting from an English speaking country. We cannot guarantee that non-natives did not take part in the study, but our subject filtering techniques based on control pairs (see Footnote 18) ensures that only the data of speakers with a good command of English were retained.

To transform binary preference data to relatedness scores about the retrieved pairs, each of them was evaluated against 50 randomly picked comparison pairs, thus it received a score on a 50-point scale (given by the number of times out of 50 the pair was picked as the most related of the two). The score was subsequently normalized between 0 and 1 by dividing the number of times the pair was picked as the most related by 50. For example, *fun-night* was chosen as more related than the comparison pair 20 times, thus its normalized score is given by $20 \div 50 = 0.4$. Note that, in each comparison, we only recorded the preference assigned to one of the two pairs, to avoid dependencies between the final scores assigned to different pairs (that is, the times a pair was selected as a random comparison item for another pair were not counted as ratings of that pair).

Because raters saw the MEN pairs matched to different random items, with the number of pairs also varying from rater to rater, it is not possible to compute annotator agreement scores for MEN. However, to get a sense of human agreement, the first and third author rated all 3,000 pairs (presented in different random orders) on a standard 1-7 Likert scale. The Spearman correlation of the two authors is at 0.68, the correlation of their average ratings with the MEN scores is at 0.84. On the one hand, this high correlation suggests that MEN contains meaningful semantic ratings. On the other, it can also be taken as an upper bound on what computational models can realistically achieve when simulating the human MEN judgments.

The high-score MEN pairs include not only pairs of terms that are strictly taxonomically close (*cathedral-church*: 0.94) but also terms that are connected by broader semantic relations, such as whole-part (*flower-petal*: 0.92), item and related event (*boat-fishing*: 0.9), etc. For this reason, we prefer to refer to MEN as a semantic *relatedness* rather than *similarity* score data set. Note that WS is also capturing a broader notion of relatedness (Agirre et al., 2009). MEN is publicly available and it can be downloaded from: `http://clic.cimec.unitn.it/~elia.bruni/MEN`.

5.2.2 Results

Table 2 reports the correlations on the MEN testing and WS data sets when using either Window2 or Window20 as textual model. Our automated tuning method selected $k = 2^9$

annotators as the most related. All control items were manually checked. Examples of control items are *hotel-word* vs. *psychology-depression*, *telephone-communication* vs. *face-locomotive*.

	Window2		Window20	
Model	*MEN*	*WS*	*MEN*	*WS*
Text	0.73	0.70	0.68	0.70
Image	0.43	0.36	0.43	0.36
NaiveFL	0.75	0.67	0.73	0.67
NaiveSL	0.76	0.69	0.74	0.64
MixLDA	0.30	0.23	0.30	0.23
Text_{mixed}	0.77	**0.73**	0.74	**0.75**
Image_{mixed}	0.55	0.52	0.57	0.51
TunedFL	**0.78**	0.72	0.76	**0.75**
TunedSL	**0.78**	0.71	**0.77**	0.72

Table 2: Spearman correlation of the models on MEN and WordSim (all coefficients significant with $p < 0.001$). TunedFL is the model selected automatically on the MEN development data; TunedSL is automatically tuned after fixing SL similarity estimation.

(when textual information comes from Window2) and $k = 2^{10}$ (with Window20) as optimal, and Feature Level (FL) similarity estimation with $\alpha = 0.5$ in both cases (since the input matrices are row-normalized, the latter setting assigns equal weights to the textual and visual components). These are the models called TunedFL in the table. The Scoring Level (SL) strategy (again with similar weights assigned to the two channels, and same k values as TunedFL) performed only slightly worse than TunedFL, and we report the results for the best SL-based models as tuned on the development MEN data as well (TunedSL). In all other models reported in the table (NaiveFL, NaiveSL, MixLDA, Text_{mixed} and Image_{mixed}), some parameters were tuned manually in order to gain insights on combination strategies representing ideas from the earlier literature.[19]

The first two rows of the table show results of the text- and image-based models, before any mixing. Text shows comparable performances on both data sets. Image correlates significantly better with MEN than WS but the correlations are lower than those of Text, in accordance with what was found in earlier studies. In the next three rows we find the results of the earlier multimodal approaches we took into consideration (Bruni et al., 2011; Feng & Lapata, 2010; Leong & Mihalcea, 2011). While the NaiveFL approach (analogous to Bruni et al.'s method), in which textual and visual matrices are concatenated without mixing, performs slightly better than Text on MEN, it attains lower performance on WS. Also NaiveSL (equivalent to Leong and Mihalcea's summing approach), where text and image sources are combined at the scoring level, obtains improvements only on MEN, loosing several correlation points on WS compared to Text.

Our implementation of MixLDA achieves very poor results both on MEN and WS. One might attribute this to the fact that Feng and Lapata's approach is constrained to using the same source for the textual and the visual model and our image data set is a poor source

19. For Text_{mixed} and Image_{mixed}, the best k values were found on the development data. They were both set to 2^{10} with both textual sources.

	Window2	Window20
Text_{mixed}	0.47	0.49
TunedFL	0.46	0.49
TunedSL	0.46	0.47

Table 3: Pearson correlation of some of our best multimodal combinations on the WordSim subset covered by Feng and Lapata (2010) (all coefficients significant with $p < 0.001$; Pearson used instead of Spearman for full comparability with Feng and Lapata). The models assigned 0 similarity to the 71/253 pairs for which they were missing a vector. Feng and Lapata (2010) report 0.32 correlation for MixLDA.

of textual data. Our approach is however also outperforming the original MixLDA by a large margin on the latter WS test set, where we are strongly disfavoured. In particular, Feng and Lapata (2010) report a correlation of 0.32 for the subset of 253 WS pairs covered by their model. We tested our system on the same subset, despite the fact that we are missing one or both vectors for 71 of the pairs (almost one third), so that our models are forced to assign 0 cosines to all these cases. Despite this huge handicap, our models are still attaining much higher correlations than the original MixLDA on the Feng and Lapata pairs, as illustrated for the most interesting fusion strategies in Table 3.

Analyzing now the effects of our fusion strategies, we can first see a uniform enhancement on both MEN and WS for Text_{mixed} and Image_{mixed} (the models obtained by first performing latent multimodal mixing on the combined matrix, but then using textual features only for Text_{mixed} and visual features only for Image_{mixed}). Text_{mixed} reaches the best performance overall on WS with both source textual models, and it is significantly better than Text on MEN according to a two-tailed paired permutation test (Moore & McCabe, 2005). Looking then at the automatically selected TunedFL model, it reaches the best performance overall. Not only it significantly outperforms Text models on both data sets, but it is significantly better than Text_{mixed} on MEN with Window20 (the difference is approaching significance with Window2 as well: $p = 0.06$). TunedSL is also very competitive. It is also significantly better than Text with both window sizes and Text_{mixed} for Window20. It is noticeably worse than TunedFL on WS with Window20 only, and it is actually having a slight advantage on MEN with Window20 (the difference between TunedFL and TunedSL is never significant).

It is worth remarking that while Text_{mixed} is a bit worse than the full fusion models, it still achieves high correlations with the human judgments and it has an extremely high correlation with the TunedFL best model ($\rho = 0.98$). This suggests that most of the benefits of multimodality are already captured by latent mixing. Text_{mixed} is an attractive model because it has less parameters than the whole pipeline and it is more compact than TunedFL, since it discards the visual features after using them for mixing.

Validating the Results While we have shown significant improvements when visual features are added to distributional models, one could object that improvements are due to the fact that we are using more information: a larger number of features (higher-dimensional vectors) for Feature Level fusion, and a more complex model (two similarity scores as independent variables to predict human judgments) for Scoring Level fusion. Further experiments provide evidence to respond to this objection.

First, we built purely textual models with the same number of features as our multimodal models – that is, instead of collecting co-occurrence of the target terms with the 30K most frequent content lemmas in our corpus (see Section 4.1 above), we extended the list of context items to the 110K most frequent content lemmas. The results with this larger textual models were virtually identical to those with 30K-dimensional vectors reported in Table 2 (correlation for the Window20 model on MEN was 0.69 instead of 0.68). Thus, at least when using our large corpus and a window-based approach, with 30K features we have pretty much exhausted the useful textual information, and it's the *nature*, not simply the quantity of the extra visual features we add that matters.

To answer the objection that the Scoring Level approach is using a more complex model, with two independent variables (text- and image-base similarities) instead of one, we casted the problem in standard inferential statistical terms (see, e.g, Baayen, 2008, ch. 6). Specifically, we fitted ordinary linear regression models to predict the MEN and WS ratings with only text-based similarities vs. text- and image-based similarities (for comparability with the Spearman correlation results reported above, the analyses were also replicated after transforming ratings and similarities into ranks). Both variables were highly significant in all experiments, and, more importantly, sequential F-tests over the nested models revealed that in all cases adding image-based similarities explains significantly more variance than what would be expected by chance given the extra parameter ($p < 0.01$).

Qualitative Analysis To acquire qualitative insights into how multimodality is contributing to meaning representation, we first picked the top 200 most related pairs from the combined MEN and WS norms, so that we would be confident that they are indeed highly related pairs for humans, and then we looked, within this subset, at those pairs with the most pronounced difference in cosines between Text and TunedFL, using Window20 as our textual source. That is, the first column of Table 4 presents pairs that are considered very related by humans and where relatedness was better captured by Text, the second column pairs where relatedness was better captured by TunedFL.

Notice that 7/10 of the relations better captured by TunedFL are between coordinates or synonyms pertaining to concrete objects (*candy/chocolate*, *bicycle/bike*, *apple/cherry*, *military/soldier*, *paws/whiskers*, *stream/waterfall* and *cheetah/lion*), that should indeed be maximally visually similar (either the objects themselves or, in a case such as *paws/whiskers*, their surrounds). The purely text-based model, on the other hand, captures relations between times of the day, that, while imageable, are not well-delimited concrete objects (*dawn/dusk*, *sunrise/sunset*). It captures properties of concepts expressed by adjectives (*dog/canine*, *skyscraper/tall*, *cat/feline*, *pregnancy/pregnant*, *rain/misty*), and at least one case where spotting the relation requires encyclopedic knowledge (*grape/wine*). We thus hypothesize that the added value of the multimodally-enhanced model derives from the power of vision in finding relations between concrete objects at the same taxonomic level, that results in detecting particularly "tight" forms of relatedness, such as synonymy and coordination.

Text	TunedFL
dawn/dusk	pet/puppy
sunrise/sunset	candy/chocolate
canine/dog	paw/pet
grape/wine	bicycle/bike
foliage/plant	apple/cherry
foliage/petal	copper/metal
skyscraper/tall	military/soldier
cat/feline	paws/whiskers
pregnancy/pregnant	stream/waterfall
misty/rain	cheetah/lion

Table 4: Top 10 pairs whose relatedness is better captured by Text (Window20) vs. TunedFL.

As observed by one reviewer, given the taxonomic nature of the information captured by the multimodal approach, it will be interesting to compare it in future work with features directly extracted from a linguistic taxonomy, such as WordNet. We observe in passing that such a manually-constructed resource, unlike those extracted from textual corpora, is likely to reflect both the linguistic and the perceptual knowledge of the lexicographers who built it.

Going in the opposite direction, another reviewer observed that we might get more mileage by combining visual features with textual models that are less taxonomic in nature. This hypothesis is partially confirmed by the fact that we obtain a larger relative improvement by mixing vision with Window20 than with Window2 (look back at Table 2, and see Section 4.1 above on why we think that the narrower window mainly captures taxonomic relations, the larger one broader topical themes). To further explore this conjecture, we re-ran the MEN and WS experiments combining the visual vectors with a document-based textual model (i.e., a semantic space whose dimensions record the number of occurrences of words in documents). Such a space is expected to capture mostly topical information, as it estimates relatedness on the basis of the tendency of words to occur in the same documents (Sahlgren, 2006). The document-based model alone was not as good as the window-based models (it obtained a Spearman correlation of 0.68 on MEN and of 0.63 on WS), and combining it with image-based models led to relative improvements comparable or inferior to those attained with Window20 (the best combined-model correlations were 0.73 on MEN and 0.70 on WS). We conclude that, while looking for textual models that are more complementary with respect to visual information seems a reasonable direction to develop multimodal systems that cover a broader range of semantic phenomena, simply emphasizing the topical side of textual models evidently does not suffice.

5.2.3 The Concreteness Factor in Modeling Relatedness Ratings: A Pilot Study

In both previous experiments, we have observed a trend towards a "division of labour" between text- and image-based models, where the latter are more apt at capturing similarity among concrete concepts and properties. One of the strongest limitations of the current version of our framework is the fact that every target word is assumed to be equally perceptually salient and consequently uniformly enriched with visual information. Intuitively, we might want to distinguish instead between concrete words, such as *chair* or *cat*, that require an integration of perceptual information for their representation, and abstract words, such as *consequence* or *absurd*, that can be represented on a purely symbolic/linguistic basis. Indeed, Recchia and Jones (2012) recently presented evidence that, in lexical decision and naming tasks, rich physical contexts favour the activation of concrete concepts, whereas rich linguistic contexts facilitate the activation of abstract concepts. With the follow-up pilot experiment presented in this section we want to pave the way for a systematic introduction of the concreteness factor in multimodal meaning representation. Operationally, we separate the abstract from the concrete word pairs in our semantic relatedness benchmark MEN, assessing the contribution of textual and visual information in approximating word meaning in the two domains independently. Importantly, we use an automated method to determine if a word is concrete or abstract, with an eye to a future integration of an automatically-determined abstractness score into our fusion algorithm.

In particular, we use abstractness scores automatically assigned by the algorithm recently introduced by Turney, Neuman, Assaf, and Cohen (2011). Scores are calculated by computing the difference between the sum of text-based semantic similarities of a target word with a set of concrete paradigm words and the sum of its semantic similarities with a set of abstract paradigm words. All words (i.e., both the paradigm words and the words for which an abstractness score was computed) were represented in a co-occurrence based matrix gathered from a large corpus of university websites. Co-occurrence counts were then transformed into Positive Pointwise Mutual Information scores (Church & Hanks, 1990) and the resulting matrix was smoothed with SVD. Pairwise semantic similarity was measured by cosines. The paradigm words were in turn selected with a supervised learning method trained on subject-rated words from the MRC Psycholinguistic Database Machine Usable Dictionary (Coltheart, 1981). Examples of highly abstract words in the automatically rated list are *purvey*: 1.00, *sense*: 0.96 and *improbable*: 0.92, while examples of highly concrete words (i.e., words with a very low abstractness score) are *donut*: 0.00, *bullet*: 0.07 and *shoe*: 0.10.

Once the abstractness score was assigned to all the MEN testing words, we divided the data set into two subsets, one containing only concrete word pairs (**MEN-conc**, 837 pairs), the other containing both abstract pairs and mixed pairs, that is pairs formed by one concrete and one abstract word (**MEN-abst**, 163 pairs). A word was considered concrete if its abstract score was ≤ 0.5, abstract otherwise. For example, the word pair *arm-bicycle* was considered concrete (with scores of 0.33 and 0.35 respectively), *fun-relax* was considered abstract (with scores of 0.6 and 0.59 respectively) and *design-orange* was considered mixed (with scores of 0.55 and 0.20 respectively). We experimented with Window20 as our purely

Model	*MEN-conc*	*MEN-abst*	*MEN-full*
Window20	0.70	0.51	0.68
Image	0.47	0.37	0.43
TunedFL	**0.78**	**0.52**	**0.76**

Table 5: Spearman correlation of the models on MEN divided into concrete and abstract subsets. Results on the full data set are also repeated. All coefficients significant with $p < 0.001$.

textual model, Image is our usual visual model and TunedFL trained on MEN development is our multimodal model.

In Table 5 we show the correlation scores for the three models on the two MEN subsets (as well as repeating the correlations they attain on the full set). First of all, it is worth noticing that all models have higher correlations with MEN-conc than MEN-abst, suggesting that approximating similarity judgments for pairs of concrete pairs is in general an easier task for distributional semantics (and, we suspect, for humans as well!). Besides this broad effect, we also observe a clear interaction for the added value of the visual component between MEN-abst and MEN-conc. In fact, TunedFL gains more than 11% in performance on MEN-conc compared to Window20, while its performance is essentially the same as that of the text-only model in the case of MEN-abst. This indicates that visual information is mostly beneficial in the concrete domain, while it maintains a neutral (timidly positive) impact on the abstract domain (recall that, in any case, MEN-abst also contains mixed pairs).

To conclude, in this section we followed up on the qualitative analysis of the main relatedness results with a pilot experiment focusing on the concreteness factor. We showed that when we divide the MEN benchmark into concrete and abstract subsets, the visual information enhances the text-based model only in the concrete domain, where its impact is very strong. We exploited an automatic scoring function to divide the data set into the concrete and abstract subsets. We can thus see the results we are reporting here also as a validation of Turney et al.'s algorithm, and, more importantly for our purposes, as an encouragement to incorporate the automated abstractness/concreteness scoring in the way in which our model mixes textual and visual information on a word-by-word basis.

5.3 Concept Categorization

To verify if the conclusions reached on WS and MEN extend to different semantic tasks and, in particular, to assess whether our multimodal approach is able to capture and organize meaning as humans do, we use two existing **concept categorization** benchmarks that we call **Battig** and Almuhareb-Poesio (**AP**), respectively, where the goal is to cluster a set of (nominal) concepts into broader categories, as already discussed in Section 2.1.

In particular, we use Battig exclusively for tuning (in the same way we used the MEN development set in the previous section) and AP for testing. Only results on AP are reported. While in the word relatedness task the tuning and testing sets were quite similar

(MEN development and MEN testing are two subsets of the same data set and the words in WS are similar to those in MEN), here the task is more challenging since Battig and AP are two independent data sets which were built following different strategies and populated with different kinds of concepts, namely very concrete and unambiguous concepts for Battig, vs. a mixture of concrete and abstract, possibly ambiguous concepts in AP. We adopted the present challenging training and testing regime because we felt that neither data set was of sufficient size to allow a split between development and testing data. More details follow.

5.3.1 Benchmarks and Method

The Battig benchmark was introduced by Baroni et al. (2010) and it is based on the Battig and Montague norms of Van Overschelde, Rawson, and Dunlosky (2004). It consists of 83 highly prototypical concepts from 10 common concrete categories (up to 10 concepts per class). Battig contains basic-level concepts belonging to categories such as *bird* (*eagle, owl...*), *kitchenware* (*bowl, spoon...*) or *vegetable* (*broccoli, potato...*). In the version we cover there are 77 concepts from 10 different classes.

AP was introduced by Almuhareb and Poesio (2005) and it is made of 402 nouns from 21 different WordNet classes. In the version we cover, AP contains 231 concepts to be clustered into 21 classes such as *vehicle* (*airplane, car...*), *time* (*aeon, future...*) or *social unit* (*brigade, nation*). The data set contains many difficult cases of unusual or ambiguous instances of a class, such as *casuarina* and *samba* as trees.

For both sets, following the original proponents and others, we cluster the words based on their pairwise cosines in the semantic space defined by a model using the CLUTO toolkit (Karypis, 2003). We use CLUTO's built-in *repeated bisections with global optimization* method, accepting all of CLUTO's default values. Cluster quality is often evaluated by percentage purity (Zhao & Karypis, 2003). If n_r^i is the number of items from the i-th true (gold standard) class that were assigned to the r-th cluster, n the total number of items, and k the number of clusters, then

$$purity = \frac{1}{\mathrm{n}} \sum_{i=1}^{i=n} \max\left(n_i^r\right)$$

In words, the number of items belonging to the majority true class (i.e., the most represented class in the cluster) are summed up across clusters and divided by the total number of items. In the best scenario purity will be 1 and it will approach 0 as cluster quality deteriorates.

Since we lack full AP coverage, the results we report below are not directly comparable with other studies that used it. However, our text-based models do have perfect coverage, and when evaluated on the full set achieve purities of 0.67 (Window2) and 0.61 (Window2), that are at state-of-the-art levels for comparable models, as reported in Section 2.1 above. So, again, we can confidently claim that the improvements achieved with multimodality are obtained by comparing our approach to competitive purely textual models.

5.3.2 Results

Table 6 reports percentage purities in the AP clustering task. Also here the best automatically selected model (TunedFL) uses FL similarity estimation as in the previous task, and has similar SVD k (2^7 for Window2 and 2^9 for Window20) and α (0.5) parameters to the

	Window2	Window20
Model	*AP*	*AP*
Text	0.73	0.65
Image	0.26	0.26
NaiveFL	0.74	0.64
NaiveSL	0.65	0.66
MixLDA	0.14	0.14
Text_{mixed}	0.74	0.67
Image_{mixed}	0.35	0.29
TunedFL	0.74	**0.69**
TunedSL	**0.75**	**0.69**

Table 6: Percentage purities of the models on AP. TunedFL is the model automatically selected on the Battig data; TunedSL is automatically tuned after fixing SL similarity estimation.

ones found for relatedness, suggesting that this particular parameter choice is robust and could be used out-of-the-box in other tasks as well. TunedSL is the best SL-based method on the tuning Battig set (same ks as TunedFL, $\alpha = 0.5$ for Window20 but $\alpha = 0.9$ on Window2).

Analogously to the previous semantic task, we see that the Image model alone is not at the level of the text models, although its AP purities are significantly above chance ($p < 0.05$ based on simulated distributions for random cluster assignment). Thus, we have a further confirmation of the fact that image-based vectors do capture important aspects of meaning. As in the previous task, MixLDA achieves very poor results.

Looking at the text-based models enhanced with visual information, we can see a general improvement in performance in almost all the multimodal combination strategies, except for NaiveFL with Window20 and NaiveSL with Window2. Even if Text_{mixed} benefits from visual smoothing in both cases, it is again outperformed by TunedFL, whose performance is here very similar to that of TunedSL, that actually is slightly better on Window2. Interestingly, TunedSL outperforms Text on Window2 despite the fact this is the single combination strongly unbalanced towards textual similarity ($\alpha = 0.9$), indicating that visual information can be beneficial even when textual information accounts for the lion's share of the composed estimate.

Like in the relatedness task, adding an equal amount of further textual features instead of image-based ones does not help with Window20 (0.66 purity with 110K textual features) and even lowers performance with Window2 (0.69 purity). Thus, the improvement brought about by visual features must be attributed to their quality, not just quantity.

According to a two-tailed permutation test, even the largest difference between TunedFL and Text on Window20 is not significant. This might be due to the brittleness of the purity statistics leading to high variance in the permutations, and possibly to suboptimal tuning. Recall, in this respect, that the tuning phase was performed on a rather different data set (Battig) compared to the data set on which we eventually evaluated the models (AP).

However, the overall trends are very encouraging, and in line with what we found in the relatedness study.

6. Conclusion

In this paper we have provided an extensive introduction to a new approach to distributional semantics that we named Multimodal Distributional Semantics. A multimodal distributional semantic model integrates a traditional text-based representation of meaning with information coming from vision. In this way, it tries to answer to the critique that distributional models lack grounding, since they base their representation of meaning entirely on the linguistic input, neglecting statistical information inherent in perceptual experience, that we humans instead exploit. Of course, a truly multimodal representation of meaning should account for the entire spectrum of human senses. On the other hand, this line of research is still in its embryonic stage and there is still a shortage of both perceptual data available and techniques to automatize their processing. This is why, in this article, we focused our analysis on the visual perceptual channel, for which we have at our disposal both large data sets and effective methods to analyze them.

In particular, we exploited the ESP-Game data set, where the image documents are tagged with words describing their content. To harvest visual information we adopted the bag-of-visual-words technique, which discretizes image content in ways that are analogous to standard text-based distributional representations. We introduced a multimodal framework that optimizes text-image fusion in a data-driven fashion on development data.

We conducted a number of experiments to assess the quality of the obtained models. We first investigated the general semantic properties of a purely image-based model, to assess its overall quality as well as to look for information complementary to that present in text. We found systematic differences between the two modalities, such as the preference for encyclopedic properties of a text-based model and for perceptual properties in the case of the image-based model. We proceeded to test a selection of models obtained by the combination of the text- and image-based representations via our multimodal framework. We used two benchmarks for word relatedness and one benchmark for word categorization and in both cases we obtained a systematic improvement in performance with the multimodal models compared to models based on standalone channels.

Still, by looking at the numerical results, we cannot deny that the improvement in performance attained when including visual information is not dramatic. Indeed, a pessimistic interpretation of the experiments could be that they confirm the hypothesis by Louwerse and others (e.g., Louwerse, 2011; Louwerse & Connell, 2011; Tillman, Datla, Hutchinson, & Louwerse, 2012) that perceptual information is already encoded, to a sufficient degree, into linguistic data, so direct visual features don't bring much to the table. However, we showed through various statistical and validation tests that our most important result, namely that adding visual information improves over using text alone, is robust and reliable. We think a more realistic take-home message is that the experiments we reported, while establishing the basic result we just mentioned, had some drawbacks we should overcome in further work.

First of all, we deliberately used general semantic benchmarks and state-of-the-art text models, so that the performance of computational methods might be getting close to the ceiling. At 0.78 correlation, our best models still have a few percentage points to go on MEN (estimated upper bound based on raters' agreement: 0.84, see Section 5.2.1), but the improvements are bound to be quite small. Concerning the AP benchmark, consider how difficult it would be even for humans to categorize *casuarina* and *samba* among the trees. Indeed, an error analysis of the TunedFL clustering results suggests that factors that might lead to better performance have little to do with vision. For example, the model "wrongly" clusters *branch* (a social unit according to AP) with the trees, and merges concepts such as *melon* and *peach* (fruit in AP) with *mandarin* and *lime* (trees). In lack of further contextual information, it's hard to dispute the model choices. Similarly, TunedFL splits the AP animal class into a cluster of small domestic mammals (*cats*, *dogs*, *kittens*, *mice*, *puppies* and *rats*) and a cluster containing everything else (mostly larger mammals such as *cows* and *elephants*). Again, the clustering procedure had no information about the classes we were searching for (e.g., animals in general, and not small animals), and so it is hard to see how performance could have improved thanks to better semantic features, visual or of other kinds. Moreover, all data sets include abstract terms, and are not specifically designed to test the more grounded aspects of meaning, where visual features might help most. We think it made sense to start our investigation with these general benchmarks of semantics, as opposed to *ad hoc* test sets, to show the viability of the multimodal approach. However, in the future we want to focus on experimental challenges where the strengths of visually-enhanced models might emerge more clearly. We took a first step in this direction by Bruni et al. (2012), where we focused specifically on how visual features can help in processing both literal and metaphorical colours.

Another factor to take into account is that both large-scale image data sets and the techniques to extract features from them are in their infancy, and we might be able to improve performance further by developing better image-based models. Regarding the data sets, we explained in Section 4.2.1 above why we chose ESP-Game, but obviously it is sub-optimal in many respects, as we also discuss there. Regarding the features, as we mentioned at the beginning of Section 4.2.2, recent advances in image processing, such as Fisher encoding, might lead to better ways to extract the information contained in images.

In the experiments, we also compared our automatically tuned multimodal model to other settings, showing its overall stability and superiority, with two important *caveats*. First, in both experiments good results are already obtained by using visual information to smooth text features, without using the visual features directly (what we called the Text_{mixed} approach). Note that this is already a multimodal approach, in that visual information is crucially used to improve the quality of the textual dimensions, and indeed we've seen that it consistently outperforms using non-multimodally-smoothed text features. While Text_{mixed} is *not* as good as our full tuned model, its simplicity makes it a very attractive approach.

Second, although automated tuning led us to prefer Feature Level over Scoring Level fusion on the development sets, TunedSL was clearly worse than TunedFL in just one case (with Window20 on WS), suggesting that, at least for the evaluation settings we considered, the difference between the two fusion strategies is not crucial. However, when comparing the "naive" versions of both strategies to the tuned ones across the results, it is clear that tuning is important to obtain consistently good performance, confirming the usefulness of our general fusion architecture.

We also conducted a pilot experiment on the concreteness/abstractness factor, to assess its impact on meaning representation and to check if it is a good candidate for a new weighted-fusion strategy we plan to investigate in the future. In fact, in the current version of the multimodal framework, the parametrization of the combination strategy works at a global level (i.e, it is the same for all words). It could be more productive to combine textual and visual information on a word-by-word basis, and tune the two modality contributions in meaning representation depending on the particular nature of each single word. Concrete vs. abstract does not constitute a neat binary distinction for all words, but it has to be rather thought as an ideal distinction to be offset with a less abrupt, real-world formulation, which takes into account the degree according to which a certain word can be considered concrete or abstract. There is no doubt that words such as *backdrop*, *squalor* or *sharp* evoke some perceptual cues gathered from our experience about them, but at the same time there is an unequivocal "amount of abstractness" accompanying them. We plan also to refine the concreteness scoring method in order to make it focus specifically on the imageable components of concreteness, as we expect them to be more relevant to our visual channel.

Further developments will focus on the techniques to extract the image-based semantic models. For example, in a pilot study (Bruni et al., 2012), we exploit new methods developed in computer vision to improve object recognition by capturing object location (Felzenszwalb, Girshick, McAllester, & Deva Ramanan, 2010; de Sande, Uijlings, Gevers, & Smeulders, 2011). We show that it is possible to extract better image-based semantic vectors by first localizing the objects denoted by words and then extracting visual information from the object location and from its surround independently. Interestingly, we discovered that image-based semantic vectors extracted from the object surround are more effective than those based on the object location when tested on our word relatedness task. For example, the fact that pictures containing deers and wolves depict similar surrounds tells us that such creatures live in similar environments, and it is thus likely that they are somewhat related. This can be seen as the distributional hypothesis transposed to images: objects that are semantically similar occur in similar visual contexts. Nevertheless, the work has to be considered a proof of concept, since we experimented with 20 words only. In future studies we will test a larger number of words.

While there is obviously much room for improvement, and many exciting routes to explore, we hope that the framework and empirical results we presented in this study convinced the reader that multimodal distributional semantics is a very promising avenue to pursue in the development of human-like models of meaning.

Acknowledgments

We thank Jasper Uijlings for his valuable suggestions about the image analysis pipeline. A lot of code and many ideas came from Giang Binh Tran, and we owe Gemma Boleda many further ideas and useful comments. Peter Turney kindly shared the abstractness score list we used in Section 5.2.3 and Yair Neuman generously helped with a preliminary analysis of the impact of abstractness on our multimodal models. Mirella Lapata kindly made the WordSim353 set used in the experiments of Feng and Lapata (2010) available to us. We thank the JAIR associated editor and reviewers for helpful suggestions and constructive criticism. Google partially funded this project with a Google Research Award to the third author. The BLESS study of Section 5.1.2 was first presented by Bruni et al. (2012).

References

Abdi, H., & Williams, L. (2010). Newman-Keuls and Tukey test. In Salkind, N., Frey, B., & Dougherty, D. (Eds.), *Encyclopedia of Research Design*, pp. 897–904. Sage, Thousand Oaks, CA.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of HLT-NAACL*, pp. 19–27, Boulder, CO.

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the web. In *Proceedings of CogSci*, pp. 103–108, Stresa, Italy.

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

Baayen, H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge, UK.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Baroni, M., Barbu, E., Murphy, B., & Poesio, M. (2010). Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, *34*(2), 222–254.

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

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

Baroni, M., & Lenci, A. (2011). How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP GEMS Workshop*, pp. 1–10, Edinburgh, UK.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Berg, T., Berg, A., & Shih, J. (2010). Automatic attribute discovery and characterization from noisy Web data. In *ECCV*, pp. 663–676, Crete, Greece.

Bergsma, S., & Goebel, R. (2011). Using visual information to predict lexical preference. In *Proceedings of RANLP*, pp. 399–405, Hissar, Bulgaria.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Image classification using random forests and ferns. In *Proceedings of ICCV*, pp. 1–8, Rio de Janeiro, Brazil.

Bosch, A., Zisserman, A., & Munoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(4).

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in Technicolor. In *Proceedings of ACL*, pp. 136–145, Jeju Island, Korea.

Bruni, E., Bordignon, U., Liska, A., Uijlings, J., & Sergienya, I. (2013). Vsem: An open library for visual semantics representation. In *Proceedings of ACL*, Sofia, Bulgaria.

Bruni, E., Tran, G. B., & Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the EMNLP GEMS Workshop*, pp. 22–32, Edinburgh, UK.

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of ACM Multimedia*, pp. 1219–1228, Nara, Japan.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13–47.

Bullinaria, J., & Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

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

Burgess, C. (2000). Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory and Language*, *43*(3), 402–408.

Caicedo, J., Ben-Abdallah, J., González, F., & Nasraoui, O. (2012). Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. *Neurocomputing*, *76*(1), 50–60.

Chatfield, K., Lempitsky, V., Vedaldi, A., & Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of BMVC*, Dundee, UK.

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clark, S. (2013). Vector space models of lexical meaning. In Lappin, S., & Fox, C. (Eds.), *Handbook of Contemporary Semantics, 2nd ed.* Blackwell, Malden, MA. In press.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*.

Connolly, A., Gleitman, L., & Thompson-Schill, S. (2007). Effect of congenital blindness on the semantic representation of some everyday concepts. *Proceedings of the National Academy of Sciences*, *104*(20), 8241–8246.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1–22, Prague, Czech Republic.

Curran, J., & Moens, M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pp. 59–66, Philadelphia, PA.

de Sande, K. V., Uijlings, J., Gevers, T., & Smeulders, A. (2011). Segmentation as selective search for object recognition. In *Proceedings of ICCV*, pp. 1879–1886, Barcelona, Spain.

de Vega, M., Glenberg, A., & Graesser, A. (Eds.). (2008). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press, Oxford, UK.

Deng, J., Dong, W., Socher, R., Li, L.-J., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pp. 248–255, Miami Beach, FL.

Dumais, S. (2003). Data-driven approaches to information access. *Cognitive Science*, *27*, 491–524.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey.. *Language and Linguistics Compass*, *6*(10), 635–653.

Escalante, H. J., Hérnadez, C. A., Sucar, L. E., & Montes, M. (2008). Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of ICMR*, Vancouver, Canada.

Evert, S. (2005). *The Statistics of Word Cooccurrences*. Dissertation, Stuttgart University.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In *Proceedings of ECCV*, Crete, Greece.

Felzenszwalb, P., Girshick, R., McAllester, D., & Deva Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 1627–1645.

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

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

Firth, J. R. (1957). *Papers in Linguistics, 1934-1951*. Oxford University Press, Oxford, UK.

Fodor, J. (1975). *The Language of Thought*. Crowell Press, New York.

Glenberg, A., & Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, *3*(43), 379–401.

Grauman, K., & Darrell, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Proceedings of ICCV*, pp. 1458–1465, Beijing, China.

Grauman, K., & Leibe, B. (2011). *Visual Object Recognition*. Morgan & Claypool, San Francisco.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer, Boston, MA.

Griffin, L., Wahab, H., & Newell, A. (2013). Distributional learning of appearance. *PLoS ONE*, *8*(2). Published online: `http://www.plosone.org/article/info:doi/10.1371/journal.pone.0058074`.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*, 1367–1368.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

Harris, Z. (1954). Distributional structure. *Word*, *10*(2-3), 1456–1162.

Johns, B., & Jones, M. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Karypis, G. (2003). CLUTO: A clustering toolkit. Tech. rep. 02-017, University of Minnesota Department of Computer Science.

Kaschak, M., Madden, C., Therriault, D., Yaxley, R., Aveyard, M., Blanchard, A., & Zwaan, R. (2005). Perception of motion affects language processing. *Cognition*, *94*, B79–B89.

Kievit-Kylar, B., & Jones, M. (2011). The Semantic Pictionary project. In *Proceedings of CogSci*, pp. 2229–2234, Austin, TX.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proceedings of CVPR*, Colorado Springs, MSA.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR*, pp. 2169–2178, Washington, DC.

Leong, C. W., & Mihalcea, R. (2011). Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of IJCNLP*, pp. 1403–1407.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–137.

Louwerse, M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Louwerse, M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*, 381–398.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pp. 1150–1157.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2).

Lowe, W. (2001). Towards a theory of semantic space. In *Proceedings of CogSci*, pp. 576–581, Edinburgh, UK.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*, 203–208.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, Cambridge, UK.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA.

McDonald, S., & Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of ACL*, pp. 17–24, Barcelona, Spain.

McRae, K., Cree, G., Seidenberg, M., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28.

Moore, D., & McCabe, G. (2005). *Introduction to the Practice of Statistics* (5 edition). Freeman, New York.

Murphy, G. (2002). *The Big Book of Concepts.* MIT Press, Cambridge, MA.

Nelson, D., McEvoy, C., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. `http://www.usf.edu/FreeAssociation/`.

Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pp. 2161–2168.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proceedings of ECCV*, pp. 490–503, Graz, Austria.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Padó, U., Padó, S., & Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP*, pp. 400–409, Prague, Czech Republic.

Pecher, D., Zeelenberg, R., & Raaijmakers, J. (1998). Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. *Journal of Memory and Language*, *38*, 401–418.

Perronnin, F., Sanchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *Proceedings of ECCV*, pp. 143–156, Berlin, Heidelberg.

Pham, T.-T., Maillot, N., Lim, J.-H., & Chevallet, J.-P. (2007). Latent semantic fusion model for image retrieval and annotation. In *Proceedings of CIKM*, pp. 439–443, Lisboa, Portugal.

Poesio, M., & Almuhareb, A. (2005). Identifying concept attributes using a classifier. In *Proceedings of the ACL Workshop on Deep Lexical Semantics*, pp. 18–27, Ann Arbor, MI.

Pulvermueller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*, 576–582.

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of WWW*, pp. 337–346, Hyderabad, India.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th MT Summit*, pp. 315–322, New Orleans, LA.

Recchia, G., & Jones, M. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, *6*(315).

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Proceedings of NAACL*, pp. 109–117, Los Angeles, CA.

Riordan, B., & 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.

Rothenhäusler, K., & Schütze, H. (2009). Unsupervised classification with dependency based word spaces. In *Proceedings of the EACL GEMS Workshop*, pp. 17–24, Athens, Greece.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Sahlgren, M. (2005). An introduction to random indexing. `http://www.sics.se/~mange/papers/RI_intro.pdf`.

Sahlgren, M. (2006). *The Word-Space Model*. Dissertation, Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*(1), 33–53.

Schütze, H. (1997). *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA.

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of EMNLP-CoNLL*, pp. 1423–1433, Jeju, Korea.

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pp. 1470–1477, Nice, France.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234–243.

Therriault, D., Yaxley, R., & Zwaan, R. (2009). The role of color diagnosticity in object recognition and representation. *Cognitive Processing*, *10*(4), 335–342.

Tillman, R., Datla, V., Hutchinson, S., & Louwerse, M. (2012). From head to toe: Embodiment through statistical linguistic frequencies. In *Proceedings of CogSci*, pp. 2434–2439, Austin, TX.

Turney, P., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pp. 680–690, Edinburgh, UK.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van de Sande, K., Gevers, T., & Snoek, C. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1582–1596.

Van Overschelde, J., Rawson, K., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289–335.

Vedaldi, A., & Fulkerson, B. (2010). Vlfeat – an open and portable library of computer vision algorithms. In *Proceedings of ACM Multimedia*, pp. 1469–1472, Firenze, Italy.

Von Ahn, L. (2006). Games with a purpose. *Computer*, *29*(6), 92–94.

Vreeswijk, D. T., Huurnink, B., & Smeulders, A. W. (2011). Text and image subject classifiers: dense works better. In *Proceedings of ACM Multimedia*, pp. 1449–1452, Scottsdale, AZ.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. In *Proceedings of CVPR*, pp. 3360–3367, San Francisco, CA.

Weeds, J. (2003). *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Department of Informatics, University of Sussex.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell, Oxford, UK. Translated by G.E.M. Anscombe.

Yang, J., Jiang, Y.-G., Hauptmann, A., & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In Wang, J. Z., Boujemaa, N., Bimbo, A. D., & Li, J. (Eds.), *Multimedia Information Retrieval*, pp. 197–206. ACM.

Zhao, Y., & Karypis, G. (2003). Criterion functions for document clustering: Experiments and analysis. Tech. rep. 01-40, University of Minnesota Department of Computer Science.