Visual Information in Semantic Representation

Yansong Feng and **Mirella Lapata**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, UK
Y.Feng-4@sms.ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

The question of how meaning might be acquired by young children and represented by adult speakers of a language is one of the most debated topics in cognitive science. Existing semantic representation models are primarily *amodal* based on information provided by the linguistic input despite ample evidence indicating that the cognitive system is also sensitive to perceptual information. In this work we exploit the vast resource of images and associated documents available on the web and develop a model of multimodal meaning representation which is based on the linguistic and visual context. Experimental results show that a closer correspondence to human data can be obtained by taking the visual modality into account.

1 Introduction

The representation and modeling of word meaning has been a central problem in cognitive science and natural language processing. Both disciplines are concerned with how semantic knowledge is acquired, organized, and ultimately used in language processing and understanding. A popular tradition of studying semantic representation has been driven by the assumption that word meaning can be learned from the linguistic environment. Words that are similar in meaning tend to behave similarly in terms of their distributions across different contexts. *Semantic space* models, among which Latent Semantic Analysis (LSA, Landauer and Dumais 1997) is perhaps known best, operationalize this idea by capturing word meaning *quantitatively* in terms of simple co-occurrence statistics. Each word w is represented by a k element vector reflecting the local distributional context of w relative to k context words. More recently, *topic models* have been gaining ground as a more structured representation of word meaning. In contrast to more standard semantic space models where word senses are conflated into a single representation, topic models assume that words observed in a corpus manifest some latent structure — word meaning is a probability distribution over a set of topics (corresponding to coarse-grained senses). Each topic is a probability distribution over words, and the content of the topic is reflected in the words to which it assigns high probability.

Semantic space (and topic) models are extracted from real language corpora, and thus provide a direct means of investigating the influence of the statistics of language on semantic representation. They have been successful in explaining a wide range of behavioral data — examples include lexical priming, deep dyslexia, text comprehension, synonym selection, and human similarity judgments (see Landauer and Dumais 1997 and the references therein). They also underlie a large number of natural language processing (NLP) tasks including lexicon acquisition, word sense discrimination, text segmentation and notably information retrieval. Despite their popularity, these models offer a somewhat impoverished representation of word meaning based solely on information provided by the linguistic input.

Many experimental studies in language acquisition suggest that word meaning arises not only from exposure to the linguistic environment but also from our interaction with the physical world. For example, infants are from an early age able to form perceptually-based category representations (Quinn et al., 1993). Perhaps unsurprisingly, words that refer to concrete entities and actions are among the first words being learned as these are directly observable in the environment (Bornstein et al., 2004). Experimental evidence also shows that children respond to categories on the basis of visual features, e.g., they generalize object names to new objects often on the basis of similarity in shape (Landau et al., 1998) and texture (Jones et al., 1991).

In this paper we aim to develop a unified mod-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 91–99,
Los Angeles, California, June 2010.

eling framework of word meaning that captures the mutual dependence between the linguistic and visual context. This is a challenging task for at least two reasons. First, in order to emulate the environment within which word meanings are acquired, we must have recourse to a corpus of verbal descriptions and their associated images. Such corpora are in short supply compared to the large volumes of solely textual data. Secondly, our model should integrate linguistic and visual information in a single representation. It is unlikely that we have separate representations for different aspects of word meaning (Rogers et al., 2004).

We meet the first challenge by exploiting multimodal corpora, namely collections of documents that contain pictures. Although large scale corpora with a one-to-one correspondence between words and images are difficult to come by, datasets that contain images and text are ubiquitous. For example, online news documents are often accompanied by pictures. Using this data, we develop a model that combines textual and visual information to learn semantic representations. We assume that images and their surrounding text have been generated by a shared set of latent variables or topics. Our model follows the general rationale of topic models — it is based upon the idea that documents are mixtures of topics. Importantly, our topics are inferred from the *joint* distribution of textual and visual words. Our experimental results show that a closer correspondence to human word similarity and association can be obtained by taking the visual modality into account.

2 Related Work

The bulk of previous work has focused on models of semantic representation that are based solely on textual data. Many of these models represent words as vectors in a high-dimensional space (e.g., Landauer and Dumais 1997), whereas probabilistic alternatives view documents as mixtures of topics, where words are represented according to their likelihood in each topic (e.g., Steyvers and Griffiths 2007). Both approaches allow for the estimation of similarity between words. Spatial models compare words using distance metrics (e.g., cosine), while probabilistic models measure similarity between terms according to the degree to which they share the same topic distributions.

Within cognitive science, the problem of how words are grounded in perceptual representations has attracted some attention. Previous modeling efforts have been relatively small-scale, using either artificial images, or data gathered from a few subjects in the lab. Furthermore, the proposed models work well for the tasks at hand (e.g., either word learning or object categorization) but are not designed as a general-purpose meaning representation. For example, Yu (2005) integrates visual information in a computational model of lexical acquisition and object categorization. The model learns a mapping between words and visual features from data provided by (four) subjects reading a children's story. In a similar vein, Roy (2002) considers the problem of learning which words or word sequences refer to objects in a synthetic image consisting of ten rectangles. Andrews et al. (2009) present a probabilistic model that incorporates perceptual information (indirectly) by combining distributional information gathered from corpus data with speaker generated feature norms[1] (which are also word-based).

Much work in computer vision attempts to learn the underlying connections between visual features and words from examples of images annotated with description keywords. The aim here is to enhance image-based applications (e.g., search or retrieval) by developing models that can label images with keywords *automatically*. Most methods discover the correlations between visual features and words by introducing latent variables. Standard latent semantic analysis (LSA) and its probabilistic variant (PLSA) have been applied to this task (Pan et al., 2004; Hofmann, 2001; Monay and Gatica-Perez, 2007). More sophisticated approaches estimate the joint distribution of words and regional image features, whilst treating annotation as a problem of statistical inference in a graphical model (Blei and Jordan, 2003; Barnard et al., 2002).

Our own work aims to develop a model of semantic representation that takes visual context into account. We do not model explicitly the correspondence of words and visual features, or learn a mapping between words and visual features. Rather, we develop a multimodal representation of meaning which is based on visual information and distributional statistics. We hypothesize that visual features are crucial in acquiring and representing meaning

[1] Participants are given a series of object names and for each object they are asked to name all the properties they can think of that are characteristic of the object.

Michelle Obama fever hits the UK

In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

Table 1: Each article in the document collection contains a document (the title is shown in boldface), and image with related content.

and conversely, that linguistic information can be useful in isolating salient visual features. Our model extracts a semantic representation from large document collections and their associated images without any human involvement. Contrary to Andrews et al. (2009) we use visual features directly without relying on speaker generated norms. Furthermore, unlike most work in image annotation, we do not employ any goldstandard data where images have been manually labeled with their description keywords.

3 Semantic Representation Model

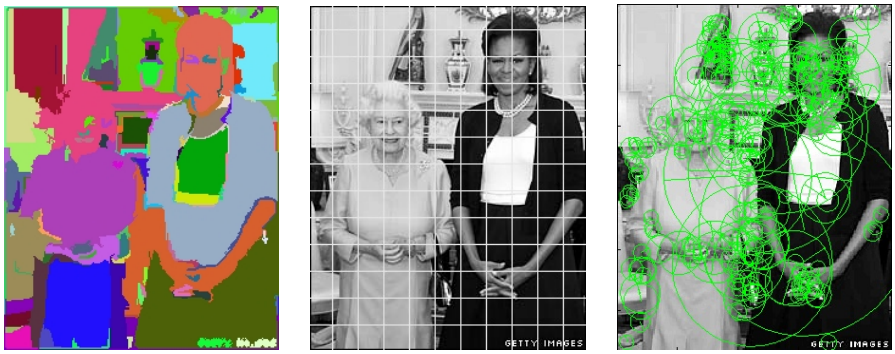
Much like LSA and the related topic models our model creates semantic representations from large document collections. Importantly, we assume that the documents are paired with images which in turn describe some of the document's content. Our experiments make use of news articles which are often accompanied with images illustrating events, objects or people mentioned in the text. Other datasets with similar properties include Wikipedia entries and their accompanying pictures, illustrated stories, and consumer photo collections. An example news article and its associated image is shown in Table 1 (we provide more detail on the database we used in our experiments in Section 4).

Our model exploits the redundancy inherent in this multimodal collection. Specifically, we assume that the images and their surrounding text have been generated by a shared set of topics. A potential stumbling block here is the fact that images and documents represent distinct modalities: images are commonly described by a continuous feature space (e.g., color, shape, texture; Barnard et al. 2002; Blei and Jordan 2003), whereas words are discrete. Fortunately, we can convert the visual features from a continuous onto a discrete space, thereby rendering image features more like word units. In the following we describe how we do this and then move on to present an extension of Latent Dirichlet Allocation (LDA, Blei and Jordan 2003), a topic model that can be used to represent meaning as a probability distribution over a set of multimodal topics. Finally, we discuss how word similarity can be measured under this model.

3.1 Image Processing

A large number of image processing techniques have been developed in computer vision for extracting meaningful features which are subsequently used in a modeling task. For example, a common first step to all automatic image annotation methods is partitioning the image into regions, using either an image segmentation algorithm (such as normalized cuts; Shi and Malik 2000) or a fixed-grid layout (Feng et al., 2004). In the first case the image is represented by irregular regions (see Figure 1(a)), whereas in the second case the image is partitioned into smaller scale regions which are uniformly extracted from a fixed grid (see Figure 1(b)). The obtained regions are further represented by a standard set of features including color, shape, and texture. These can be treated as continuous vectors (Blei and Jordan, 2003) or in quantized form (Barnard et al., 2002).

Despite much progress in image segmentation, there is currently no automatic algorithm that can reliably divide an image into meaningful parts. Extracting features from small local regions is thus preferable, especially for image collections that are diverse and have low resolution (this is often the case for news images). In our work we identify local regions using a difference-of-Gaussians point detector (see Figure 1(c)). This representation is based on descriptors computed over automatically detected image regions. It provides a much richer (and hopefully more informative) feature space compared to the alternative image representations discussed above. For example, an image segmentation algorithm, would extract at most 20 regions from the image in Figure 1; uniform grid segmentation yields 143



(a) (b) (c)

Figure 1: Image partitioned into regions of varying granularity using (a) the normalized cut image segmentation algorithm, (b) uniform grid segmentation, and (c) the SIFT point detector.

(11×13) regions, whereas an average of 240 points (depending on the image content) are detected. A non-sparse feature representation is critical in our case, since we usually do not have more than one image per document.

We compute local image descriptors using the the Scale Invariant Feature Transform (SIFT) algorithm (Lowe, 1999). Importantly, SIFT descriptors are designed to be invariant to small shifts in position, changes in illumination, noise, and viewpoint and can be used to perform reliable matching between different views of an object or scene (Mikolajczyk and Schmid, 2003; Lowe, 1999). We further quantize the SIFT descriptors using the K-means clustering algorithm to obtain a discrete set of visual terms (visiterms) which form our visual vocabulary Voc_V. Each entry in this vocabulary stands for a group of image regions which are similar in content or appearance and assumed to originate from similar objects. More formally, each image I is expressed in a bag-of-words format vector, $[v_1, v_2, ..., v_L]$, where $v_i = n$ only if I has n regions labeled with v_i. Since both images and documents in our corpus are now represented as bags-of-words, and since we assume that the visual and textual modalities express the same content, we can go a step further and represent the document and its associated image as a mixture of verbal and visual words d_{Mix}. We will then learn a topic model on this concatenated representation of visual and textual information.

3.2 Topic Model

Latent Dirichlet Allocation (Blei et al., 2003; Griffiths et al., 2007) is a probabilistic model of text generation. LDA models each document using a mixture over K topics, which are in turn characterized as distributions over words. The words in the document are generated by repeatedly sampling a topic according to the topic distribution, and selecting a word given the chosen topic. Under this framework, the problem of meaning representation is expressed as one of statistical inference: given some data — textual and visual words — infer the latent structure from which it was generated. Word meaning is thus modeled as a probability distribution over a set of latent multimodal topics.

LDA can be represented as a three level hierarchical Bayesian model. Given a corpus consisting of M documents, the generative process for a document d is as follows. We first draw the mixing proportion over topics θ_d from a Dirichlet prior with parameters α. Next, for each of the N_d words w_{dn} in document d, a topic z_{dn} is first drawn from a multinomial distribution with parameters θ_{dn}. The probability of a word token w taking on value i given that topic $z = j$ is parametrized using a matrix β with $b_{ij} = p(w = i|z = j)$. Integrating out θ_d's and z_{dn}'s, gives $P(D|\alpha,\beta)$, the probability of a corpus (or document collection):

$$\prod_{d=1}^{M} \int P(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

The central computational problem in topic modeling is to compute the posterior distribution $P(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)$ of the hidden variables given a document $\mathbf{w} = (w_1, w_2, \ldots, w_N)$. Although this distribution is intractable in general, a variety of ap-

proximate inference algorithms have been proposed in the literature including variational inference which our model adopts. Blei et al. (2003) introduce a set of variational parameters, γ and ϕ, and show that a tight lower bound on the log likelihood of the probability can be found using the following optimization procedure:

$$(\gamma^*, \phi^*) = \underset{\gamma,\phi}{\operatorname{argmin}} D(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

Here, D denotes the Kullback-Leibler (KL) divergence between the true posterior and the variational distribution $q(\theta, \mathbf{z}|\gamma, \phi)$ defined as: $q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)$, where the Dirichlet parameter γ and the multinomial parameters $(\phi_1, \ldots, \phi_N)$ are the free variational parameters. Notice that the optimization of parameters $(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w}))$ is document-specific (whereas α is corpus specific).

Previous applications of LDA (e.g., to document classification or information retrieval) typically make use of the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ associated with a given document. We are not so much interested in γ as we wish to obtain a semantic representation for a given word across documents. We therefore train the LDA model sketched above on a corpus of multimodal documents $\{d_{Mix}\}$ consisting of both textual and visual words. We select the number of topics, K, and apply the LDA algorithm to obtain the β parameters, where β represents the probability of a word w_i given a topic z_j, $p(w_i|z_j) = \beta_{ij}$. The meaning of w_i is thus extracted from β and is a K-element vector, whose components correspond to the probability of w_i given each latent topic assumed to have generated the document collection.

3.3 Similarity Measures

The ability to accurately measure the similarity or association between two words is often used as a diagnostic for the psychological validity of semantic representation models. In the topic model described above, the similarity between two words w_1 and w_2 can be intuitively measured by the extent to which they share the same topics (Griffiths et al., 2007). For example, we may use the KL divergence to measure the difference between the distributions p and q:

$$D(p,q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j}$$

where p and q are shorthand for $P(w_1|z_j)$ and $P(w_2|z_j)$, respectively.

The KL divergence is asymmetric and in many applications, it is preferable to apply a symmetric measure such as the Jensen Shannon (JS) divergence. The latter measures the "distance" between p and q through $\frac{(p+q)}{2}$, the average of p and q:

$$JS(p,q) = \frac{1}{2}\left[D(p, \frac{(p+q)}{2}) + D(q, \frac{(p+q)}{2})\right]$$

An alternative approach to expressing the similarity between two words is proposed in Griffiths et al. (2007). The underlying idea is that word association can be expressed as a conditional distribution. If we have seen word w_1, then we can determine the probability that w_2 will be also generated by computing $P(w_2|w_1)$. Although the LDA generative model allows documents to contain multiple topics, here it is assumed that both w_1 and w_2 came from a single topic:

$$P(w_2|w_1) = \sum_{z=1}^{K} P(w_2|z)P(z|w_1)$$

$$P(z|w_1) \propto P(w_1|z)P(z)$$

where $p(z)$ is uniform, a single topic is sampled from the distribution $P(z|w_1)$, and an overall estimate is obtained by averaging over all topics K.

Griffiths et al. (2007) report results on modeling human association norms using exclusively $P(w_2|w_1)$. We are not aware of any previous work that empirically assesses which measure is best at capturing semantic similarity. We undertake such an empirical comparison as it is not a priory obvious how similarity is best modeled under a multimodal representation.

4 Experimental Setup

In this section we discuss our experimental design for assessing the performance of the model presented above. We give details on our training procedure and parameter estimation and present the baseline method used for comparison with our model.

Data We trained the multimodal topic model on the corpus created in Feng and Lapata (2008). It contains 3,361 documents that have been downloaded from the BBC News website.[2] Each document comes with an image that depicts some of its content. The images are usually 203 pixels wide

[2] http://news.bbc.co.uk/

and 152 pixels high. The average document length is 133.85 words. The corpus has 542,414 words in total. Our experiments used a vocabulary of 6,253 textual words. These were words that occurred at least five times in the whole corpus, excluding stopwords. The accompanying images were preprocessed as follows. We first extracted SIFT features from each image (150 on average) which we subsequently quantized into a discrete set of visual terms using K-means. As we explain below, we determined an optimal value for K experimentally.

Evaluation Our evaluation experiments compared the multimodal topic model against a standard text-based topic model trained on the same corpus whilst ignoring the images. Both models were assessed on two related tasks, that have been previously used to evaluate semantic representation models, namely word association and word similarity.

In order to simulate word association, we used the human norms collected by Nelson et al. (1999).[3] These were established by presenting a large number of participants with a cue word (e.g., *rice*) and asking them to name an associate word in response (e.g., *Chinese, wedding, food, white*). For each word, the norms provide a set of associates and the frequencies with which they were named. We can thus compute the probability distribution over associates for each cue. Analogously, we can estimate the degree of similarity between a cue and its associates using our model (and any of the measures in Section 3.3). And consequently examine (using correlation analysis) the degree of linear relationship between the human cue-associate probabilities and the automatically derived similarity values. We also report how many times the word with the highest probability under the model was the first associate in the norms. The norms contain 10,127 unique words in total. Of these, we created semantic representations for the 3,895 words that appeared in our corpus.

Our word similarity experiment used the WordSim353 test collection (Finkelstein et al., 2002) which consists of relatedness judgments for word pairs. For each pair, a similarity judgment (on a scale of 0 to 10) was elicited from human subjects (e.g., *tiger-cat* are very similar, whereas *delay–racism* are not). The average rating for each pair represents an estimate of the perceived similarity of the two words. The task varies slightly from word association. Here, participants are asked to rate perceived similarity rather than generate the first word that came into their head in response to a cue word. The collection contains similarity ratings for 353 word pairs. Of these, we constructed semantic representations for the 254 that appeared in our corpus. We also evaluated how well model produced similarities correlate with human ratings. Throughout this paper we report correlation coefficients using Pearson's r.

[3] http://www.usf.edu/Freeassociation.

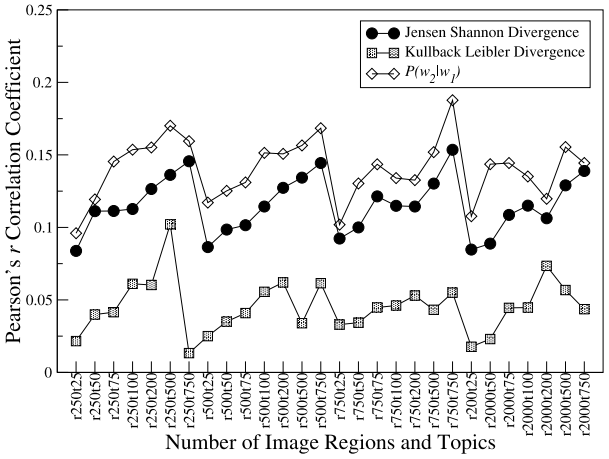


Figure 2: Performance of multimodal topic model on predicting word association under varying topics and visual terms (development set).

5 Experimental Results

Model Selection The multimodal topic model has several parameters that must be instantiated. These include the quantization of the image features, the number of topics, the choice of similarity function, and the values for α and β. We explored the parameter space on held-out data. Specifically, we fit the parameters for the word association and similarity models separately using a third of the association norms and WordSim353 similarity judgments, respectively. As mentioned in Section 3.1 we used K-means to quantize the image features into a discrete set of visual terms. We varied K from 250 to 2000. We also varied the number of topics from 25 to 750 for both the multimodal and text-based topic models. The parameter α was set to 0.1 and β was initialized randomly. The model was trained using variational Bayes until convergence of its bound on the likelihood objective. This took 1,000 iterations.

Figure 2 shows how word association performance varies on the development set with different numbers of topics (t) and visual terms (r) according

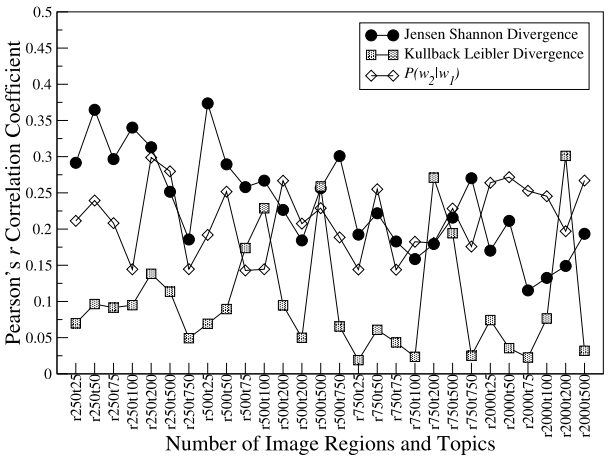


Figure 3: Performance of multimodal topic model on predicting word similarity under varying topics and visual terms (development set).

to three similarity measures: KL divergence, JS divergence, and $P(w_2|w_1)$, the probability of word w_2 given w_1 (see Section 3.3). Figure 3 shows results on the development set for the word similarity task. As far as word association is concerned, we obtain best results with $P(w_2|w_1)$, 750 visual terms and 750 topics ($r = 0.188$). On word similarity, JS performs best with 500 visual terms and 25 topics ($r = 0.374$). It is not surprising that $P(w_2|w_1)$ works best for word association. The measure expresses the associative relations between words as a conditional distribution over potential response words w_2 for cue word w_1. A symmetric function is more appropriate for word similarity as the task involves measuring the degree to which to words share some meaning (expressed as topics in our model) rather than whether a word is likely to be generated as a response to another word. These differences also lead to different parametrizations of the semantic space. A rich visual term vocabulary (750 terms) is needed for modeling association as broader aspects of word meaning are taken into account, whereas a sparser more focused representation (with 500 visual terms and 25 overall topics) is better at isolating the common semantic content between two words. We explored the parameter space for the text-based topic model in a similar fashion. On the word association task the best correlation coefficient was achieved with 750 topics and $P(w_2|w_1)$ ($r = 0.139$). On word similarity, the best results were obtained with 75 topics and the JS divergence ($r = 0.309$).

Model	Word Association	Word Similarity
UpperBnd	0.400	0.545
MixLDA	0.123	0.318
TxtLDA	0.077	0.247

Table 2: Model performance on word association and similarity (test set).

Model Comparison Table 2 summarizes our results on the test set using the optimal set of parameters as established on the development set. The first row shows how well humans agree with each other on the two tasks (UpperBnd). We estimated the intersubject correlation using leave-one-out resampling[4] (Weiss and Kulikowski, 1991). As can be seen, in all cases the topic model based on textual and visual modalities (MixLDA) outperforms the model relying solely on textual information (TxtLDA). The differences in performance are statistically significant ($p < 0.05$) using a t-test (Cohen and Cohen, 1983).

Steyvers and Griffiths (2007) also predict word association using Nelson's norms and a state-of-the-art LDA model. Although they do not report correlations, they compute how many times the word with the highest probability $P(w_2|w_1)$ under the model was the first associate in the human norms. Using a considerably larger corpus (37,651 documents), they reach an accuracy of 16.15%. Our corpus contains 3,361 documents, the MixLDA model performs at 14.15% and the LDA model at 13.16%. Using a vector-based model trained on the BNC corpus (100M words), Washtell and Markert (2009) report a correlation of 0.167 on the same association data set, whereas our model achieves a correlation of 0.123. With respect to word similarity, Marton et al. (2009) report correlations within the range of 0.31–0.54 using different instantiations of a vector-based model trained on the BNC with a vocabulary of 33,000 words. Our MixLDA model obtains a correlation of 0.318 with a vocabulary five times smaller (6,253 words). Although these results are not strictly comparable due to the different nature and size of the training data, they give some indication of the quality of our model in the context of other approaches that exploit only the textual modality. Besides, our intent is not to report the best performance possible,

[4] We correlated the data obtained from each participant with the ratings obtained from all other participants and report the average.

GAME, CONSOLE, XBOX, SECOND, SONY, WORLD, TIME, JAPAN, JAPANESE, SCHUMACHER, LAP, MICROSOFT, ALONSO, RACE, TITLE, WIN, GAMERS, LAUNCH, RENAULT, MARKET
PARTY, MINISTER, BLAIR, LABOUR, PRIME, LEADER, GOVERNMENT, TELL, BROW, MP, TONY, SIR, SECRETARY, ELECTION, CONFERENCE, POLICY, NEW, WANT, PUBLIC, SPEECH
SCHOOL, CHILD, EDUCATION, STUDENT, WORK, PUPIL, PARENT, TEACHER, GOVERNMENT, YOUNG, SKILL, AGE, NEED, UNIVERSITY, REPORT, LEVEL, GOOD, HELL, NEW, SURVEY

Table 3: Most frequent words in three topics learnt from a corpus of image-document pairs.

but to show that a model of meaning representation is more accurate when taking visual information into account.

Table 3 shows some examples of the topics found by our model, which largely form coherent blocks of semantically related words. In general, we observe that the model using image features tends to prefer words that visualize easily (e.g., CONSOLE, XBOX). Furthermore, the visual modality helps obtain crisper meaning distinctions. Here, SCHUMACHER is a very probable world for the "game" cluster. This is because the Formula One driver appears as a character in several video games discussed and depicted in our corpus. For comparison the "game" cluster for the text-based LDA model contains the words: GAME, USE, INTERNET, SITE, USE, SET, ONLINE, WEB, NETWORK, MURRAY, PLAY, MATCH, GOOD, WAY, BREAK, TECHNOLOGY, WORK, NEW, TIME, SECOND.

We believe the model presented here works better than a vanilla text-based topic model for at least three reasons: (1) the visual information helps create better clusters (i.e., conceptual representations) which in turn are used to measure similarity or association; these clusters themselves are amodal but express commonalities across the visual and textual modalities; (2) the model is also able to capture perceptual correlations between words. For example, RED is the most frequent associate for APPLE in Nelson's norms. This association is captured in our visual features (pictures with apples cluster with pictures showing red objects) even though RED does not co-occur with APPLE in our data; (3) finally, even in cases where two words are visually very different in terms of shape or color (e.g., BANANA and APPLE), they tend to appear in images with similar structure (e.g., on tables, in bowls, as being held or eaten by someone) and thus often share some common element of meaning.

6 Conclusion

In this paper we developed a computational model that unifies visual and linguistic representations of word meaning. The model learns from natural language corpora paired with images under the assumption that visual terms and words are generated by mixtures of latent topics. We have shown that a closer correspondence to human data can be obtained by explicitly taking the visual modality into account in comparison to a model that estimates the topic structure solely from the textual modality. Beyond word similarity and association, the approach is promising for modeling word learning and categorization as well as a wide range of priming studies. Outwith cognitive science, we hope that some of the work described here might be of relevance to more applied tasks such as thesaurus acquisition, word sense disambiguation, multimodal search, image retrieval, and summarization.

Future improvements include developing a non-parametric version that jointly *learns* how many visual terms and topics are optimal. Currently, the size of the visual vocabulary and the number of topics are parameters in the model, that must be tuned separately for different tasks and corpora. Another extension concerns the creation of visual terms. Our model assumes that an image is a bag of words. The assumption is convenient for modeling purposes, but clearly false in the context of visual processing. Image descriptors found closely to each other are likely to represent the same object and should form one term rather than several distinct ones (Wang and Grimson, 2007). Taking the spatial structure among visual words into account would yield better topics and overall better semantic representations. Analogously, we could represent documents by their syntactic structure (Boyd-Graber and Blei, 2009).

References

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

Barnard, K., P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. 2002. Matching words and pic-

tures. *Journal of Machine Learning Research* 3:1107–1135.

Blei, D. and M. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference*. Toronto, ON, pages 127–134.

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

Bornstein, M. H., L. R. Cote, S. Maital, K. Painter, S.-Y. Park, and L. Pascual. 2004. Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development* 75(4):1115–1139.

Boyd-Graber, J. and D. Blei. 2009. Syntactic topic models. In *Proceedings of the 22nd Conference on Advances in Neural Information Processing Systems*. MIT, Press, Cambridge, MA, pages 185–192.

Cohen, J. and P. Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Feng, S., V. Lavrenko, and R. Manmatha. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*. Washington, DC, pages 1002–1009.

Feng, Y. and M. Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceedings of the ACL-08: HLT*. Columbus, pages 272–280.

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

Griffiths, T. L., M. Steyvers, and J. B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review* 114(2):211–244.

Hofmann, T. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 41(2):177–196.

Jones, S. S., L. B. Smith, and B. Landau. 1991. Object properties and knowledge in early lexical learning. *Child Development* (62):499–516.

Landau, B., L. Smith, and S. Jones. 1998. Object perception and object naming in early development. *Trends in Cognitive Science* 27:19–24.

Landauer, T. and S. T. Dumais. 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.

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

Marton, Y., S. Mohammad, and P. Resnik. 2009. Estimating semantic distance using soft semantic constraints in knowledge-source – corpus hybrid models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, pages 775–783.

Mikolajczyk, K. and C. Schmid. 2003. A performance evaluation of local descriptors. In *Proceedings of the 9th International Conference on Computer Vision and Pattern Recognition*. Nice, France, volume 2, pages 257–263.

Monay, F. and D. Gatica-Perez. 2007. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10):1802–1817.

Nelson, D. L., C. L. McEvoy, and T.A. Schreiber. 1999. The university of South Florida word association norms.

Pan, J., H. Yang, P. Duygulu, and C. Faloutsos. 2004. Automatic image captioning. In *Proceedings of the 2004 International Conference on Multimedia and Expo*. Taipei, pages 1987–1990.

Quinn, P., P. Eimas, and S. Rosenkrantz. 1993. Evidence for representations of perceptually similar natural categories by 3-month and 4-month old infants. *Perception* 22:463–375.

Rogers, T. T., M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, and K. Patterson. 2004. Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review* 111(1):205–235.

Roy, D. 2002. Learning words and syntax for a visual description task. *Computer Speech and Language* 16(3).

Shi, J. and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Steyvers, M. and T. Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S Dennis, and W Kintsch, editors, *A Handbook of Latent Semantic Analysis*, Psychology Press.

Wang, X. and E. Grimson. 2007. Spatial latent Dirichlet allocation. In *Proceedings of the 20th Conference on Advances in Neural Information Processing Systems*. MI Press, Cambridge, MA, pages 1577–1584.

Washtell, J. and K. Markert. 2009. A comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore, pages 628–637.

Weiss, S. M. and C. A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann, San Mateo, CA.

Yu, C. 2005. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science* 17(3):381–397.