A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities

Stephen Roller
Department of Computer Science
The University of Texas at Austin
roller@cs.utexas.edu

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
schulte@ims.uni-stuttgart.de

Abstract

Recent investigations into grounded models of language have shown that holistic views of language and perception can provide higher performance than independent views. In this work, we improve a two-dimensional multimodal version of Latent Dirichlet Allocation (Andrews et al., 2009) in various ways. (1) We outperform text-only models in two different evaluations, and demonstrate that low-level visual features are directly compatible with the existing model. (2) We present a novel way to integrate visual features into the LDA model using unsupervised clusters of images. The clusters are directly interpretable and improve on our evaluation tasks. (3) We provide two novel ways to extend the bimodal models to support three or more modalities. We find that the three-, four-, and five-dimensional models significantly outperform models using only one or two modalities, and that nontextual modalities each provide separate, disjoint knowledge that cannot be forced into a shared, latent structure.

1 Introduction

In recent years, an increasing body of work has been devoted to multimodal or "grounded" models of language where semantic representations of words are extended to include perceptual information. The underlying hypothesis is that the meanings of words are explicitly tied to our perception and understanding of the world around us, and textual-information alone is insufficient for a complete understanding of language.

The *language grounding problem* has come in many different flavors with just as many different approaches. Some approaches apply semantic parsing, where words and sentences are mapped to logical structure meaning (Kate and Mooney, 2007). Others provide automatic mappings of natural language instructions to executable actions, such as interpreting navigation directions (Chen and Mooney, 2011) or robot commands (Tellex et al., 2011; Matuszek et al., 2012). Some efforts have tackled tasks such as automatic image caption generation (Feng and Lapata, 2010a; Ordonez et al., 2011), text illustration (Joshi et al., 2006), or automatic location identification of Twitter users (Eisenstein et al., 2010; Wing and Baldridge, 2011; Roller et al., 2012).

Another line of research approaches grounded language knowledge by augmenting distributional approaches of word meaning with perceptual information (Andrews et al., 2009; Steyvers, 2010; Feng and Lapata, 2010b; Bruni et al., 2011; Silberer and Lapata, 2012; Johns and Jones, 2012; Bruni et al., 2012a; Bruni et al., 2012b; Silberer et al., 2013). Although these approaches have differed in model definition, the general goal in this line of research has been to enhance word meaning with perceptual information in order to address one of the most common criticisms of distributional semantics: that the "meaning of words is entirely given by other words" (Bruni et al., 2012b).

In this paper, we explore various ways to integrate new perceptual information through novel computational modeling of this grounded knowledge into a multimodal distributional model of word meaning. The model we rely on was originally developed by

Andrews et al. (2009) and is based on a generalization of Latent Dirichlet Allocation. This model has previously been shown to provide excellent performance on multiple tasks, including prediction of association norms, word substitution errors, semantic inferences, and word similarity (Andrews et al., 2009; Silberer and Lapata, 2012). While prior work has used the model only with feature norms and visual attributes, we show that low-level image features are directly compatible with the model and provide improved representations of word meaning. We also show how simple, unsupervised clusters of images can act as a semantically useful and qualitatively interesting set of features. Finally, we describe two ways to extend the model by incorporating three or more modalities. We find that each modality provides useful but disjoint information for describing word meaning, and that a hybrid integration of multiple modalities provides significant improvements in the representations of word meaning. We release both our code and data to the community for future research.[1]

2 Related Work

The language grounding problem has received significant attention in recent years, owed in part to the wide availability of data sets (e.g. Flickr, Von Ahn (2006)), computing power, improved computer vision models (Oliva and Torralba, 2001; Lowe, 2004; Farhadi et al., 2009; Parikh and Grauman, 2011) and neurological evidence of ties between the language, perceptual and motor systems in the brain (Pulvermüller et al., 2005; Tettamanti et al., 2005; Aziz-Zadeh et al., 2006).

Many approaches to multimodal research have succeeded by abstracting away raw perceptual information and using high-level representations instead. Some works abstract perception via the usage of symbolic logic representations (Chen et al., 2010; Chen and Mooney, 2011; Matuszek et al., 2012; Artzi and Zettlemoyer, 2013), while others choose to employ concepts elicited from psycholinguistic and cognition studies. Within the latter category, the two most common representations have been association norms, where subjects are given a cue word and name the first (or several) associated words that come to mind (e.g., Nelson et al. (2004)), and feature norms, where subjects are given a cue word and asked to describe typical properties of the cue concept (e.g., McRae et al. (2005)).

Griffiths et al. (2007) helped pave the path for cognitive-linguistic multimodal research, showing that Latent Dirichlet Allocation outperformed Latent Semantic Analysis (Deerwester et al., 1990) in the prediction of association norms. Andrews et al. (2009) furthered this work by showing that a bimodal topic model, consisting of both text and feature norms, outperformed models using only one modality on the prediction of association norms, word substitution errors, and semantic interference tasks. In a similar vein, Steyvers (2010) showed that a different feature-topic model improved predictions on a fill-in-the-blank task. Johns and Jones (2012) take an entirely different approach by showing that one can successfully infer held out feature norms from weighted mixtures based on textual similarity. Silberer and Lapata (2012) introduce a new method of multimodal integration based on Canonical Correlation Analysis, and performs a systematic comparison between their CCA-based model and others on association norm prediction, held out feature prediction, and word similarity.

As computer vision techniques have improved over the past decade, other research has begun directly using visual information in place of feature norms. The first work to do this with topic models is Feng and Lapata (2010b). They use a Bag of Visual Words (BoVW) model (Lowe, 2004) to create a bimodal vocabulary describing documents. The topic model using the bimodal vocabulary outperforms a purely textual based model in word association and word similarity prediction. Bruni et al. (2012a) show how a BoVW model may be easily combined with a distributional vector space model of language using only vector concatenation. Bruni et al. (2012b) show that the *contextual* visual words (i.e. the visual features around an object, rather than of the object itself) are even more useful at times, suggesting the plausibility of a sort of distributional hypothesis for images. More recently, Silberer et al. (2013) show that visual attribute classifiers, which have been immensely successful in object recognition (Farhadi et al., 2009), act as excellent substitutes for feature

[1] `http://stephenroller.com/research/emnlp13`

norms. Other work on modeling the meanings of verbs using video recognition has also begun showing great promise (Mathe et al., 2008; Regneri et al., 2013).

The Computer Vision community has also benefited greatly from efforts to unify the two modalities. To name a few examples, Rohrbach et al. (2010) and Socher et al. (2013) show how semantic information from text can be used to improve zero-shot classification (i.e., classifying never-before-seen objects), and Motwani and Mooney (2012) show that verb clusters can be used to improve activity recognition in videos.

3 Data

Our experiments use several existing and new data sets for each of our modalities. We employ a large web corpus and a large set of association norms. We also introduce two new overlapping data sets: a collection of feature norms and a collection of images for a number of German nouns.

3.1 Textual Modality

For our **Text** modality, we use deWaC, a large German web corpus created by the WaCKy group (Baroni et al., 2009) containing approximately 1.7B word tokens. We filtered the corpus by: removing words with unprintable characters or encoding troubles; removing all stopwords; removing word types with a total frequency of less than 500; and removing documents with a length shorter than 100. The resulting corpus has 1,038,883 documents consisting of 75,678 word types and 466M word tokens.

3.2 Cognitive Modalities

Association Norms (AN) is a collection of association norms collected by Schulte im Walde et al. (2012). In association norm experiments, subjects are presented with a cue word and asked to list the first few words that come to mind. With enough subjects and responses, association norms can provide a common and detailed view of the meaning components of cue words. After removing responses given only once in the entire study, the data set contains a total of 95,214 cue-response pairs for 1,012 nouns and 5,716 response types.

Feature Norms (FN) is our new collection of feature norms for a group of 569 German nouns. We present subjects on Amazon Mechanical Turk with a cue noun and ask them to give between 4 and 8 typical descriptive features of the noun. Subjects are given ten example responses; one such example is a cue of *Tisch* 'table' and a response of *hat Beine* 'has legs'. After collection, subjects who are obvious spammers or did not follow instructions are manually filtered. Responses are manually corrected for spelling mistakes and semantically normalized.[2] Finally, responses which are only given once in the study are removed. The final data set contains 11,714 cue-response pairs for 569 nouns and 2,589 response types.

Note that the difference between association norms and feature norms is subtle, but important. In AN collection, subjects simply name related words as fast as possible, while in FN collection, subjects must carefully *describe* the cue.

3.3 Visual Modalities

BilderNetle ("little ImageNet" in Swabian German) is our new data set of German noun-to-ImageNet synset mappings. ImageNet is a large-scale and widely used image database, built on top of WordNet, which maps words into groups of images, called synsets (Deng et al., 2009). Multiple synsets exist for each meaning of a word. For example, ImageNet contains two different synsets for the word *mouse*: one contains images of the animal, while the other contains images of the computer peripheral. This BilderNetle data set provides mappings from German noun types to images of the nouns via ImageNet.

Starting with a set of noun compounds and their nominal constituents von der Heide and Borgwaldt (2009), five native German speakers and one native English speaker (including the authors of this paper) work together to map German nouns to ImageNet synsets. With the assistance of a German-English dictionary, the participants annotate each word with all its possible meanings. After discussing the annotations with the German speakers, the English speaker manually map the word meanings to synset senses in ImageNet. Finally, the German speakers review samples of the images for each word to en-

[2] For brevity, we include the full details of the spammer identification, cleansing process and normalization techniques in the Supplementary Materials.

sure the pictures accurately reflect the original noun in question. Not all words or meanings are mapped to ImageNet, as there are a number of words without entries in ImageNet, but the resulting data set contains a considerable amount of polysemy. The final data set contains 2022 word-synset mappings for just 309 words. All but three of these words overlap with our data set of feature norms. After extracting sections of images using bounding boxes when available by ImageNet (and using the entire image when bounding boxes are unavailable), the data set contains 1,305,602 images.

3.3.1 Image Processing

After the collection of all the images, we extracted simple, low-level computer vision features to use as modalities in our experiments.

First, we compute a simple Bag of Visual Words (BoVW) model for our images using SURF keypoints (Bay et al., 2008). SURF is a method for selecting points-of-interest within an image. It is faster and more forgiving than the commonly known SIFT algorithm. We compute SURF keypoints for every image in our data set using SimpleCV[3] and randomly sample 1% of the keypoints. The keypoints are clustered into 5,000 visual codewords (centroids) using k-means clustering (Sculley, 2010), and images are then quantized over the 5,000 codewords. All images for a given word are summed together to provide an average representation for the word. We refer to this representation as the **SURF** modality.

While this is a standard, basic BoVW model, each individual codeword on its own may not provide a large degree of semantic information; typically a BoVW representation acts predominantly as a feature space for a classifier, and objects can only be recognize using collections of codewords. To test that similar concepts should share similar visual codewords, we cluster the BoVW representations for all our images into 500 clusters with k-means clustering, and represent each word as membership over the image clusters, forming the **SURF Clusters** modality. The number of clusters is chosen arbitrarily. Ideally, each cluster should have a common object or clear visual attribute, and words are express in terms of these visual commonalities.

We also compute GIST vectors (Oliva and Torralba, 2001) for every image using LearGIST (Douze et al., 2009). Unlike SURF descriptors, GIST produces a single vector representation for an image. The vector does not find points of interest in the image, but rather attempts to provide a representation for the overall “gist” of the whole image. It is frequently used in tasks like scene identification, and Deselaers and Ferrari (2011) shows that distance in GIST space correlates well with semantic distance in WordNet. After computing the GIST vectors, each textual word is represented as the centroid GIST vector of all its images, forming the **GIST** modality.

Finally, as with the SURF features, we clustered the GIST representations for our images into 500 clusters, and represented words as membership in the clusters, forming the **GIST Clusters** modality.

4 Model Definition

Our experiments are based on the multimodal extension of Latent Dirichlet Allocation developed by Andrews et al. (2009). Previously LDA has been successfully used to infer unsupervised joint topic distributions over words and feature norms together (Andrews et al., 2009; Silberer and Lapata, 2012). It has also been shown to be useful in joint inference of text with visual attributes obtained using visual classifiers (Silberer et al., 2013). These multimodal LDA models (hereafter, mLDA) have been shown to be qualitatively sensible and highly predictive of several psycholinguistic tasks (Andrews et al., 2009). However, prior work using mLDA is limited to two modalities at a time. In this section, we describe bimodal mLDA and define two methods for extending it to three or more modalities.

4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei et al., 2003), or LDA, is an unsupervised Bayesian probabilistic model of text documents. It assumes that all documents are probabilistically generated from a shared set of K common topics, where each topic is a multinomial distribution over the vocabulary (notated as β), and documents are modeled as mixtures of these shared topics (notated as θ). LDA assumes every document in the corpus is generated using the fol-

[3] http://simplecv.org

lowing generative process:

1. A document-specific topic distribution, $\theta_d \sim Dir(\alpha)$ is drawn.
2. For the ith word in the document,
   (a) A topic assignment $z_i \sim \theta_d$ is drawn,
   (b) and a word $w_i \sim \beta_{z_i}$ is drawn and observed.

The task of Latent Dirichlet Allocation is then to automatically infer the latent document distribution θ_d for each document $d \in \mathcal{D}$, and the topic distribution β_k for each of the $k = \{1, \ldots, K\}$ topics, given the data. The probability that the ith word of document d is

$$p(w_i, \theta_d) = \sum_k p(w_i|\beta_k)p(z_i = k|\theta_d).$$

4.2 Multimodal LDA

Andrews et al. (2009) extend LDA to allow for the inference of document and topic distributions in a multimodal corpus. In their model, a document consists of a set of (word, feature) pairs,[4] rather than just words, and documents are still modeled as mixtures of shared topics. Topics consist of multinomial distributions over words, β_k, but are extended to also include multinomial distributions over features, ψ_k. The generative process is amended to include these feature distributions:

1. A document-specific topic distribution, $\theta_d \sim Dir(\alpha)$ is drawn.
2. For the ith (word, feature) pair in the document,
   (a) A topic assignment $z_i \sim \theta_d$ is drawn;
   (b) a word $w_i \sim \beta_{z_i}$ is drawn;
   (c) a feature $f_i \sim \psi_{z_i}$ is drawn;
   (d) the pair (w_i, f_i) is observed.

The conditional probability of the ith pair (w_i, f_i) is updated appropriately:

$$p(w_i, f_i, \theta_d) = \sum_k p(w_i|\beta_k)p(f_i|\psi_k)p(z_i = k|\theta_d).$$

The key aspect to notice is that the observed word w_i and feature f_i are conditionally independent given the topic selection, z_i. This powerful extension allows for joint inference over both words and features, and topics become the key link between the text and feature modalities.

4.3 3D Multimodal LDA

We can easily extend the bimodal LDA model to incorporate three or more modalities by simply performing inference over n-tuples instead of pairs, and still mandating that each modality is conditionally independent given the topic. We consider the ith tuple $(w_i, f_i, f'_i, \ldots)$ in document d to have a conditional probability of:

$$p(w_i, f_i, f'_i, \ldots, \theta_d) = \sum_k p(w_i|\beta_i)p(f_i|\psi_k)p(f'_i|\psi'_i) \cdots p(z_i = k|\theta_d)$$

That is, we simply take the original mLDA model of Andrews et al. (2009) and generalize it in the same way they generalize LDA. At first glance, it seems that the inference task should become more difficult as the number of modalities increases and observed tuples become sparser, but the task remains roughly the same difficulty, as all of the observed elements of a tuple are conditionally independent given the topic assignment z_i.

4.4 Hybrid Multimodal LDA

3D Multimodal LDA assumes that all modalities share the same latent topic structure, θ_d. It is possible, however, that all modalities do *not* share some latent structure, but the modalities can still combine in order to enhance word meaning. The intuition here is that language usage is guided by all information gained in all modalities, but knowledge gained from one modality may not always relate to another modality. For example, the color red and the feature "is sweet" both enhance our understanding of strawberries. However, one cannot see that strawberries are sweet, so one should not correlate the color red with the feature "is sweet."

To this end, we define Hybrid Multimodal LDA. In this setting, we perform separate, bimodal mLDA inference according to Section 4.2 for each of the different modalities, and then concatenate the topic distributions for the words. In this way, Hybrid mLDA assumes that every modality shares some latent structure with the text in the corpus, but the latent structures are not shared between non-textual modalities.

[4] Here, and elsewhere, *feature* and f simply refer to a token from a nontextual modality and should not be confused with the machine learning sense of *feature*.

For example, to generate a hybrid model for text, feature norms and SURF, we separately perform bimodal mLDA for the text/feature norms modalities and the text/SURF modalities. This provides us with two topic-word distributions: $\beta^{FN}_{k,w}$ and $\beta^{S}_{k',w}$, and the hybrid model is simply the concatenation of the two distributions,

$$\beta^{FN\&S}_{j,w} = \begin{cases} \beta^{FN}_{j,w} & 1 \leq j \leq K^{FN}, \\ \beta^{S}_{j-K^{FN},w} & K^{FN} < j \leq K^{FN} + K^{S}, \end{cases}$$

where K^{FN} indicates the number of topics for the Feature Norm modality, and likewise for K^{S}.

4.5 Inference

Analytical inference of the posterior distribution of mLDA is intractable, and must be approximated. Prior work using mLDA has used Gibbs Sampling to approximate the posterior, but we found this method did not scale with larger values of K, especially when applied to the relatively large deWaC corpus.

To solve these scaling issues, we implement Online Variational Bayesian Inference (Hoffman et al., 2010; Hoffman et al., 2012) for our models. In Variational Bayesian Inference (VBI), one approximates the true posterior using simpler distributions with free variables. The free variables are then optimized in an EM-like algorithm to minimize difference between the true and approximate posteriors. Online VBI differs from normal VBI by using randomly sampled minibatches in each EM step rather than the entire data set. Online VBI easily scales and quickly converges in all of our experiments. A listing of the inference algorithm may be found in the Supplementary Materials and the source code is available as open source.

5 Experimental Setup

5.1 Generating Multimodal Corpora

In order to evaluate our algorithms, we first need to generate multimodal corpora for each of our non-textual modalities. We use the same method as Andrews et al. (2009) for generating our multimodal corpora: for each word token in the text corpus, a feature is selected stochastically from the word's feature distribution, creating a word-feature pair. Words without grounded features are all given the same *placeholder* feature, also resulting in a word-feature pair.[5] That is, for the feature norm modality, we generate (word, feature norm) pairs; for the SURF modality, we generate (word, codeword) pairs, etc. The resulting stochastically generated corpus is used in its corresponding experiments.

The 3D text-feature-association norm corpus is generated slightly differently: for each word in the original text corpus, we check the existence of multimodal features in either modality. If a word had no features, it is represented as a triple (word, placeholder$_{FN}$, placeholder$_{AN}$). If the word had only feature norms, but no associations, it is generated as (word, feature, placeholder$_{AN}$), and similarly for association norms without feature norms. In the case of words with presence in both modalities, we generate *two* triples: (word, feature, placeholder$_{AN}$) and (word, placeholder$_{FN}$, association). This allows association norms and feature norms to influence each other via the document mixtures θ, but avoids falsely labeling explicit relationships between randomly selected feature norms and associations.[6] Other 3D corpora are generated using the same general procedure.

5.2 Evaluation

We evaluate each of our models with two data sets: a set of compositionality ratings for a number of German noun-noun compounds, and the same association norm data set used as one of our training modalities in some settings.

Compositionality Ratings is a data set of compositionality ratings originally collected by von der Heide and Borgwaldt (2009). The data set consists of 450 concrete, depictable German noun compounds along with compositionality ratings with regard to their constituents. For each compound, 30 native German speakers are asked to rate how related the meaning of the compound is to each of its constituents on a scale from 1 (highly opaque; entirely noncompositional) to 7 (highly transparent; very compositional). The mean of the 30 judgments

[5] Placeholder features must be hardcoded to have equal probability over all topics to prevent all placeholder pairs from aggregating into a single topic.

[6] We did try generating the random triples without placeholders, but the generated explicit relationships are overwhelmingly detrimental in the settings we attempted.

is taken as the gold compositionality rating for each of the compound-constituent pairs. For example, *Ahornblatt* 'maple leaf' is rated highly transparent with respect to its constituents, *Ahorn* 'maple' and *Blatt* 'leaf', but *Löwenzahn* 'dandelion' is rated non-compositional with respect to its constituents, *Löwe* 'lion' and *Zahn* 'tooth'.

We use a subset of the original data, comprising of all two-part noun-noun compounds and their constituents. This data set consists of 488 compositionality ratings (244 compound-head and 244 compound-modifier ratings) for 571 words. 309 of the targets have images (the entire image data set); 563 have feature norms; and all 571 of have association norms.

In order to predict compositionality, for each compound-constituent pair $(w_{compound}, w_{constituent})$, we compute negative symmetric KL divergence between the two words' topic distributions, where symmetric KL divergence is defined as

$$sKL(w_1||w_2) = KL(w_1||w_2) + KL(w_2||w_1),$$

and KL divergence is defined as

$$KL(w_1||w_2) = \sum_k \ln\left(\frac{p(t=k|w_1)}{p(t=k|w_2)}\right) p(t=k|w_1).$$

The values of $-sKL$ for all compound-constituent word pairs are correlated with the human judgments of compositionality using Spearman's ρ, a rank-order correlation coefficient. Note that, since KL divergence is a measure of *dissimilarity*, we use *negative* symmetric KL divergence so that our ρ correlation coefficient is positive. For example, we compute both $-sKL(Ahornblatt, Ahorn)$ and $-sKL(Ahornblatt, Blatt)$, and so on for all 488 compound-constituent pairs, and then correlate these values with the human judgments.

Additionally, we also evaluate using the **Association Norms** data set described in Section 3. Since it is not sensible to evaluate association norm prediction when they are also used as training data, we omit this evaluation for this modality. Following Andrews et al. (2009), we measure association norm prediction as an average of percentile ranks. For all possible pairs of words in our vocabulary, we compute the negative symmetric KL divergence between the two words. We then compute the percentile ranks of similarity for each word pair, e.g., "cat" is more similar to "dog" than 97.3% of the rest of the vocabulary. We report the weighted mean percentile ranks for all cue-association pairs, i.e., if a cue-association is given more than once, it is counted more than once.

5.3 Model Selection and Hyperparameter Optimization

In all settings, we fix all Dirichlet priors at 0.1, use a learning rate 0.7, and use minibatch sizes of 1024 documents. We do not optimize these hyperparameters or vary them over time. The high Dirichlet priors are chosen to prevent sparsity in topic distributions, while the other parameters are selected as the best from Hoffman et al. (2010).

In order to optimize the number of topics K, we run five trials of each modality for 2000 iterations for $K = \{50, 100, 150, 200, 250\}$ (a total of 25 runs per setup). We select the value or K for each model which minimizes the average perplexity estimate over the five trials.

6 Results

6.1 Predicting Compositionality Ratings

Table 1 shows our results for each of our selected models with our compositionality evaluation. The 2D models employing feature norms and association norms do significantly better than the text-only model (two-tailed t-test). This result is consistent with other works using this model with these features (Andrews et al., 2009; Silberer and Lapata, 2012).

We also see that the SURF visual words are able to provide notable, albeit not significant, improvements over the text-only modality. This confirms that the low-level BoVW features do carry semantic information, and are useful to consider individually. The GIST vectors, on the other hand, perform almost exactly the same as the text-only model. These features, which are usually more useful for comparing overall image likeness than object likeness, do not *individually* contain semantic information useful for compositionality prediction.

The performance of the visual modalities reverses when we look at our cluster-based models. Text

Modality	K	ρ
Text Only		
Text Only (LDA)	200	.204
Bimodal mLDA		
Text + Feature Norms	150	.310 ***
Text + Assoc. Norms	200	.328 **
Text + SURF	50	.251
Text + GIST	100	.204
Text + SURF Clusters	200	.159
Text + GIST Clusters	150	.233
3D mLDA		
Text + FN + AN	250	.259
Text + FN + SURF	100	.286 *
Text + FN + GC	200	.261 *
Hybrid mLDA		
FN, AN	150+200	.390 ***
FN, SURF	150+50	.350 ***
FN, GC	150+150	.340 ***
FN, AN, GC	150+200+150	.395 ***
FN, AN, SURF	150+200+50	.404 ***
FN, AN, SURF, GC	150+200+50+150	**.406** ***

Table 1: Average rank correlations between $-sKL(w_{compound}, w_{constituent})$ and our Compositionality gold standard. The Hybrid models are the concatenation of the corresponding Bimodal mLDA models. Stars indicate statistical significance compared to the text-only setting at the .05, .01 and .001 levels using a two-tailed t-test.

Modality	K	Assoc.
Text Only		
Text Only (LDA)	200	.679
Bimodal mLDA		
Text + Feature Norms	150	.676
Text + SURF	50	.789 ***
Text + GIST	100	.739 ***
Text + SURF Clusters	200	.618 ***
Text + GIST Clusters	150	.690
3D mLDA		
Text + FN + SURF	100	.722 ***
Text + FN + GC	200	.601 ***
Hybrid mLDA		
FN, SURF	150+50	.800 ***
FN, GC	150+150	.742 ***
FN, GC, SURF	150+150+50	**.804** ***

Table 2: Average predicted rank similarity between cue words and their associates. Stars indicate statistical significance compared to the text-only modality, with gray stars indicating the model is statistically worse than the text model. The Hybrid models are the concatenation of the corresponding Bimodal mLDA models.

combined with SURF clusters is our worst performing system, indicating our clusters of images with common visual words are actively working against us. The clusters based on GIST, on the other hand, provide a minor improvement in compositionality prediction.

All of our 3D models are better than the text-only model, but they show a performance drop relative to one or both of their comparable bimodal models. The model combining text, feature norms, and association norms is especially surprising: despite the excellent performance of each of the bimodal models, the 3D model performs significantly worse than either of its components ($p < .05$). This indicates that these modalities provide new insight into word meaning, but cannot be forced into the same latent structure.

The hybrid models show massive performance increases across the board. Indeed, our 5 modality hybrid model obtains a performance nearly twice that of the text-only model. Not only do all 6 hybrid models do significantly better than the text-only models, they show a highly significant improvement over their individual components ($p < .001$ for all 16 comparisons). Furthermore, improvements generally continue to grow significantly with each additional modality we incorporate into the hybrid model ($p < .001$ for all but the .404 to .406 comparison, which is not significant). Clearly, there is a great deal to learn from combining three, four and even five modalities, but the modalities are learning *disjoint* knowledge which cannot be forced into a shared, latent structure.

6.2 Predicting Association Norms

Table 2 shows the average weighted predicted rank similarity between all cue words and associates and trials. Here we see that feature norms do not seem to be improving performance on the association norms. This is slightly unexpected, but consistent with the result that feature norms seem to provide helpful, but disjoint semantic information as association norms.

We see that the image modalities are much more useful than they are in compositionality prediction. The SURF modality does extremely well in particular, but the GIST features also provide statistically significant improvements over the text-only model. Since the SURF and GIST image features tend to capture object-likeness and scene-likeness respectively, it is possible that words which share associates are likely related through common settings and objects that appear with them. This seems to provide additional evidence of Bruni et al. (2012b)'s suggestion that something like a distributional hypothesis of images is plausible.

Once again, the clusters of images using SURF causes a dramatic drop in performance. Combined with the evidence from the compositionality assessment, this shows that the SURF clusters are actively confusing the models and not providing semantic information. GIST clusters, on the other hand, are providing a marginal improvement over the text-only model, but the result is not significant. We take a qualitative look into the GIST clusters in the next section.

Once again, we see that the 3D models are ineffective compared to their bimodal components, but the hybrid models provide at least as much information as their components. The Feature Norms and GIST Clusters hybrid model significantly improves over both components.[7] The final four-modality hybrid significantly outperforms all comparable models. As with the compositionality evaluation, we conclude that the image and and feature norm models are providing disjoint semantic information that cannot be forced into a shared latent structure, but still augment each other when combined.

7 Qualitative Analysis of Image Clusters

In all research connecting word meaning with perceptual information, it is desirable that the inferred representations be directly interpretable. One nice property of the cluster-based modalities is that we may represent each cluster as its prototypical images, and examine whether the prototypes are related to the topics.

We chose to limit our analysis to the GIST clusters for two primary reasons: first, the SURF clusters did not perform well in our evaluations, and second, preliminary investigation into the SURF clusters show that the majority of SURF clusters are nearly identical. This indicates our SURF clusters are likely hindered by poor initialization or parameter selection, and may partially explain their poor performance in evaluations.

We select our single best Text + GIST Clusters trial from the Compositionality evaluation and look at the topic distributions for words and image clusters. For each topic, we select the three clusters with the highest weight for the topic, $p(c|\psi_k)$. We extract the five images closest to the cluster centroids, and select two topics whose prototypical images are the most interesting and informative. Figure 1 shows these selected topics.

The first example topic contains almost exclusively water-related terms. The first image, extracted from the most probable cluster, does not at first seem related to water. Upon further inspection, we find that many of the water-related pictures are scenic views of lakes and mountains, often containing a cloudy sky. It seems that the GIST cluster does not tend to group images of water, but rather nature scenes that may contain water. This relationship is more obvious in the second picture, especially when one considers the water itself contains reflections of the trees and mountain.

The second topic contains time-related terms. The "@card@" term is a special token for all non-zero and non-one numbers. The second word, "Uhr", is polysemous: it can mean *clock*, an object which tells the time, or *o'clock*, as in *We meet at 2 o'clock* ("Wir treffen uns um 2 Uhr.") The three prototypical pictures are not pictures of clocks, but round, detailed objects similar to clocks. We see GIST has a preference toward clustering images based on the predominant shape of the image. Here we see the clusters of GIST images are not providing a definite semantic relationship, but an overwhelming visual one.

8 Conclusions

In this paper, we evaluated the role of low-level image features, SURF and GIST, for their compatibility with the multimodal Latent Dirichlet Allocation model of Andrews et al. (2009). We found both fea-

[7] The gain is smaller than compared to SURF Hybrid, but there is much less variance in the trials.

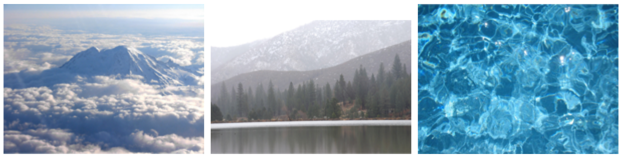
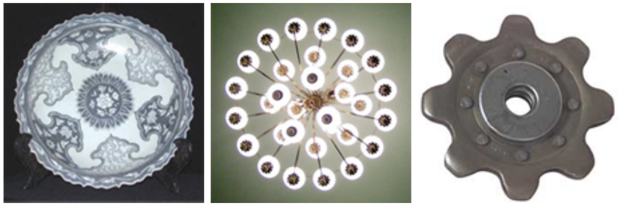
Most Probable Words	Translations	Prototypical Images
Wasser Schiff See Meer Meter Fluß	water ship lake sea meter river	
@card@ Uhr Freitag Sonntag Samstag Montag	(number) clock Friday Sunday Saturday Monday	

Figure 1: Example topics with prototypical images for the Text + GIST Cluster modality. The first topic shows water-related words, as well scenes which often appear with water. The second shows clock-like objects, but not clocks.

ture sets were directly compatible with multimodal LDA and provided significant gains in their ability to predict association norms over traditional text-only LDA. SURF features also provided significant gains over text-only LDA in predicting the compositionality of noun compounds.

We also showed that words may be represented in terms of membership of image clusters based on the low-level image features. Image clusters based on GIST features were qualitatively interesting, and were able to give improvements over the text-only model.

Finally, we showed two methods for extending multimodal LDA to three or more modalities: the first as a 3D model with a shared latent structure between all modalities, and the second where latent structures were inferred separately for each modality and joined together into a hybrid model. Although the 3D model was unable to compete with its bimodal components, we found the hybrid model consistently improved performance over its component modalities. We conclude that the combination of many modalities provides the best representation of word meaning, and that each nontextual modality is discovering disjoint information about word meaning that cannot be forced into a global latent structure.

Acknowledgments

We would like to thank the UT Natural Language Learning Reading Group and the anonymous EMNLP reviewers for their most helpful comments and suggestions. We would also like to thank the members of the SemRel group at IMS for their considerable help in the construction of BilderNetle.

The authors acknowledge the Texas Advanced Computing Center (TACC) for providing the grid computing resources necessary for these results. The research presented in this paper was funded by the DFG Collaborative Research Centre SFB 732 (Stephen Roller) and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. In *Transactions of the Association for Computational Linguistics*, volume 1, pages 49–62.

Lisa Aziz-Zadeh, Stephen M. Wilson, Giacomo Rizzolatti, and Marco Iacoboni. 2006. Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology*, 16(18):1818–1823.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed webcrawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, June.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional semantics from text and images. *Proceedings of the EMNLP 2011 Geometrical Models for Natural Language Semantics*, pages 22–32.

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

Elia Bruni, Jasper Uijlings, Marco Baroni, and Nicu Sebe. 2012b. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 1219–1228.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pages 859–865, August.

David L. Chen, Joohyun Kim, and Raymond J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37(1):397–436.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Thomas Deselaers and Vittorio Ferrari. 2011. Visual and semantic similarity in imagenet. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1777–1784.

Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. 2009. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 19:1–19:8.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785.

Yansong Feng and Mirella Lapata. 2010a. How many words is a picture worth? Automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249.

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

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211.

Matthew Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23:856–864.

Matthew Hoffman, David M. Blei, Chong Wang, and John Paisley. 2012. Stochastic variational inference. *ArXiv e-prints*, June.

Brendan T. Johns and Michael N. Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120.

Dhiraj Joshi, James Z. Wang, and Jia Li. 2006. The story picturing engine—a system for automatic text illustration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):68–89.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambiguous supervision. In *Proceedings of the 22nd Conference on Artificial Intelligence*, volume 7, pages 895–900.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Stefan Mathe, Afsaneh Fazly, Sven Dickinson, and Suzanne Stevenson. 2008. Learning the abstract motion semantics of verbs from captioned videos. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2012. Learning to parse natural language commands to a robot control system. In *Proceedings of the 13th International Symposium on Experimental Robotics*.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Tanvi S. Motwani and Raymond J. Mooney. 2012. Improving video activity recognition using object recognition and text mining. In *ECAI*, pages 600–605.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *International Conference on Computer Vision*, pages 503–510. IEEE.

Friedemann Pulvermüller, Olaf Hauk, Vadim V. Nikulin, and Risto J Ilmoniemi. 2005. Functional links between motor and language systems. *European Journal of Neuroscience*, 21(3):793–797.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. In *Transactions of the Association for Computational Linguistics*, volume 1, pages 25–36.

Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. 2010. What helps where–and why? Semantic relatedness for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 910–917.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510.

Sabine Schulte im Walde, Susanne Borgwaldt, and Ronny Jauch. 2012. Association norms of german noun compounds. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 632–639, Istanbul, Turkey.

D. Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1177–1178.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea, July.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August.

Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-shot learning through cross-modal transfer. *International Conference on Learning Representations*.

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

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth J. Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

Marco Tettamanti, Giovanni Buccino, Maria Cristina Saccuman, Vittorio Gallese, Massimo Danna, Paola Scifo, Ferruccio Fazio, Giacomo Rizzolatti, Stefano F. Cappa, and Daniela Perani. 2005. Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, 17(2):273–281.

Luis Von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 11, pages 955–964.