Grounded Models of Semantic Representation

Carina Silberer and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
c.silberer@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

A popular tradition of studying semantic representation has been driven by the assumption that word meaning can be learned from the linguistic environment, despite ample evidence suggesting that language is *grounded* in perception and action. In this paper we present a comparative study of models that represent word meaning based on linguistic and perceptual data. Linguistic information is approximated by naturally occurring corpora and sensorimotor experience by feature norms (i.e., attributes native speakers consider important in describing the meaning of a word). The models differ in terms of the mechanisms by which they integrate the two modalities. Experimental results show that a closer correspondence to human data can be obtained by uncovering latent information shared among the textual and perceptual modalities rather than arriving at semantic knowledge by concatenating the two.

1 Introduction

Distributional models of lexical semantics have seen considerable success at accounting for a wide range of behavioral data in tasks involving semantic cognition (Landauer and Dumais, 1997; Griffiths et al., 2007). These models have also enjoyed lasting popularity in natural language processing. Examples involve information retrieval (Salton et al., 1975), word sense discrimination (Schütze, 1998), text segmentation (Choi et al., 2001), and numerous studies of lexicon acquisition (Grefenstette, 1994; Lin, 1998). Despite their widespread use, distributional models have been criticized as "disembodied" in that they learn exclusively from linguistic information but are not *grounded* in perception and action (Perfetti, 1998; Barsalou, 1999; Glenberg and Kaschak, 2002).

This lack of grounding contrasts with many experimental studies suggesting that word meaning is acquired not only from exposure to the linguistic environment but also from our interaction with the physical world (Landau et al., 1998; Bornstein et al., 2004). Beyond language acquisition, there is considerable evidence across both behavioral experiments and neuroimaging studies that the perceptual associates of words play an important role in language processing (for a review see Barsalou (2008)).

It is thus no surprise that recent years have witnessed the emergence of perceptually grounded distributional models. An important question in the formulation of such models concerns the provenance of perceptual information. A few models use feature norms as a proxy for sensorimotor experience (Howell et al., 2005; Andrews et al., 2009; Steyvers, 2010; Johns and Jones, 2012). These are obtained by asking native speakers to write down attributes they consider important in describing the meaning of a word. The attributes represent perceived physical and functional properties associated with the referents of words. For example, *apples* are typically green or red, round, shiny, smooth, crunchy, tasty, and so on; *dogs* have four legs and bark, whereas *chairs* are used for sitting. Other models focus solely on the visual modality under the assumption that it represents a major source of data from

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1423–1433, Jeju Island, Korea, 12–14 July 2012.

which humans can learn semantic representations of both linguistic and non-linguistic communicative actions (Regier, 1996). For example, Feng and Lapata (2010) learn semantic representations from corpora of texts paired with naturally co-occurring images (e.g., news articles and their associated pictures), whereas Bruni et al. (2011) learn textual and visual representations independently from distinct data sources.

Aside from the type of data used to capture perceptual information, another important issue concerns how the two modalities (perceptual and textual) are integrated. A simple solution would be to learn both modalities independently (Bruni et al., 2011) or to infer one modality by means of the other (Johns and Jones, 2012) and to arrive at a grounded representation simply by concatenating the two. An alternative is to learn from both modalities *jointly* (Andrews et al., 2009; Feng and Lapata, 2010; Steyvers, 2010). According to this view, semantic knowledge is gained by simultaneously learning from the statistical structure within each modality assuming both data sources have been generated by a shared set of meanings or topics.

In this paper we undertake the first comparative study of perceptually grounded distributional models. We examine three models with different assumptions regarding the integration of perceptual and linguistic data. The first model, originally proposed by Andrews et al. (2009), is an extension of latent Dirichlet allocation (LDA, Blei et al. (2003)). It simultaneously considers the distribution of words across contexts in a text corpus and the distribution of words across perceptual features and extracts joint information from both data sources. Our second model is based on Johns and Jones (2012) who represent the meaning of a word as the concatenation of its textual and its perceptual vector. Interestingly, their model allows to infer a perceptual vector for words without feature norms, simply by taking into account similar words for which perceptual information is available.

Finally, we propose Canonical Correlation Analysis (Hotelling, 1936; Hardoon et al., 2004) as our third model. CCA is a data analysis and dimensionality reduction method similar to PCA. While PCA deals with only one data space, CCA is a technique for joint dimensionality reduction across two (or more) spaces that provide heterogeneous representations of the same objects. The assumption is that the representations in these two spaces contain some joint information that is reflected in correlations between them.

Features	table	dog	apple
has_4_legs	.28	.60	0
used_for_eating	.50	0	0
a_pet	0	.40	0
is_brown	0	0	0
is_crunchy	0	0	.58
is_round	.22	0	.42
has_fangs	0	0	0

Table 1: Feature norms for the nouns *table*, *dog*, and *apple* shown as a distribution.

In all three models we use feature norms as a proxy for perceptual information. Despite their shortcomings (e.g., they often cover a small fraction of the vocabulary of an adult speaker due to the effort involved in eliciting them), feature norms provide detailed knowledge about meaning representations and are a useful starting point for studying the integration of perceptual and textual information without being susceptible to the effects of noise, e.g., coming from image processing. In other words, feature norms can serve as an upper bound of what can be achieved when integrating detailed perceptual information with vanilla text-based distributional models.

Our experimental results demonstrate that joint models give a better fit to human word similarity and association data than a model that considers only one data source, or the simple concatenation of the two sources.

2 Perceptually Grounded Models

In this study we examine semantic representation models that rely on linguistic and perceptual data. The linguistic environment is approximated by corpora such as the British National Corpus (BNC). As mentioned earlier, we resort to feature norms as proxy for perceptual information. In our experiments, we relied on the norming study of McRae et al. (2005), in which a large number of human participants were presented with a series of words and

asked to list relevant features of the words' referents. Table 1 presents examples of features participants listed for the nouns *apple*, *dog*, and *table*. The number of participants listing a certain feature for a word can be used to compute a probability distribution over features given the word:

$$P(f_k|w) = \frac{f(f_k,w)}{\sum_{m=1}^{F} f(f_m,w)} \quad (1)$$

where $f(f_k,w)$ is the number of participants who listed feature f_k for word w and F is the total number of features.

In the remainder of this section we will describe our models and how they arrive at an integrated perceptual and linguistic representation.

2.1 Feature-topic Model

Andrews et al. (2009) present an extension of LDA (Blei et al., 2003) where words in documents as well as their associated features are treated as observed variables that are explained by a generative process. The underlying training data consists of a corpus $\mathcal{D}$ where each document is represented by words and their frequency of occurrence within the document. In addition, those words of a document that are also included in the feature norms are paired with one of their features, where a feature is sampled according to the feature distribution given that word. For example, suppose a document d_j consists of the sentence *Mix in the apple, celery, raisins, and apple juice.* Suppose further that all content words except of *mix* and *juice* are included in the feature norms. Then, a representation for d_j is *mix*:1, *apple*;is_red:2, *celery*;has_leaves:1, *raisin*;is_edible:1, *juice*:1.

The plate diagram in Figure 1 illustrates the graphical model in detail. Each document d_j in $\mathcal{D}$ is generated by a mixture of components $\{x_1,...,x_c,...,x_C\} \in \mathcal{C}$; a component x_c comprises a latent discourse topic coupled with a feature cluster originating from the feature norms. A discourse topic belonging to x_c, in turn, is a distribution $\phi_c \in \phi = \{\phi_1,...,\phi_C\}$ over words, and a feature cluster is a distribution $\psi_c \in \psi = \{\psi_1,...,\psi_C\}$ over features.

In order to create document d_j, a distribution π_j over components is sampled from a Dirichlet distribution parametrized by α. To generate each word $w_{ji} \in \{w_{j1},...,w_{jn_j}\}$, a component $x_c = x_{ji}$ is drawn from π_j; w_{ji} is then drawn from the corresponding distribution ϕ_c. If w_{ji} is in the feature norms, it is coupled with a feature f_{ji} which is correspondingly drawn from ψ_c. A symmetric Dirichlet prior with hyperparameters β and γ is placed on ϕ and ψ, respectively. The probability of the corpus $\mathcal{D}$ is defined as:

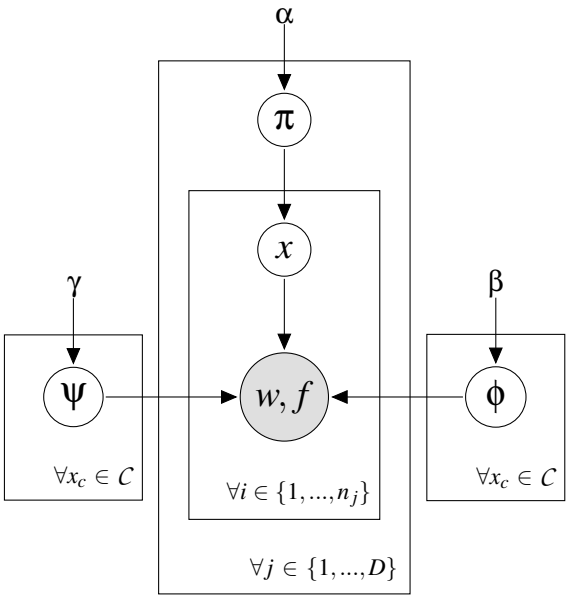

Figure 1: Feature-topic model. The components x_{ji} of a document d_j are sampled from π_j. For each $x_c = x_{ji}$, a word w_{ji} is drawn from distribution ϕ_c and a feature f_{ji} is drawn from distribution ψ_c.

$$P((w \cup f)_{1:D}|\phi,\psi,\alpha) = \prod_{j=1}^{D} \int d\pi_j \prod_{i=1}^{n_j} P(\pi_j|\alpha) \sum_{c=1}^{C} P(w_{ji}|x_{ji}=x_c,\phi) P(f_{ji}|x_{ji}=x_c,\psi) P(x_{ji}=x_c|\pi_j) \quad (2)$$

where D is the number of documents and C the predefined number of components. Computing the posterior distribution $P(\phi,\psi,\alpha,\beta,\gamma|(w \cup f)_{1:D})$ of the hidden variables given the data is generally intractable:

$$P(\phi,\psi,\alpha,\beta,\gamma|(w \cup f)_{1:D}) \propto P((w \cup f)_{1:D}|\phi,\psi,\alpha) P(\phi|\beta)P(\psi|\gamma)P(\alpha)P(\beta)P(\gamma) \quad (3)$$

Equation (3) may be approximated using the Gibbs

	x_1	x_2	x_{12}	...	x_{28}	x_{75}	x_{107}	x_{119}	x_{125}	x_{148}	x_{182}	...	x_{266}	x_{326}	x_{349}	x_{350}
apple	3e-5	3e-5	0	...	5e-4	9e-4	.09	.002	7.6e-5	2e-4	.003	...	0	0	3e-6	0

Figure 2: Example of the representation of the meaning of *apple* with the model of (Andrews et al., 2009) .

	...	d_{16}	...	d_{322}	...	d_{2469}	d_{2470}	...	d_D	a_fruit	has_fangs	$is\_crunchy$	...	is_yellow	is_red	is_green	is_round
apple	...	1	...	1	...	0	1	...	0	0	0	0	...	0	0	0	0
apple	...	1	...	1	...	0	1	...	0	.006	1.8e-5	8e-4	...	.004	.004	.006	.02

Figure 3: Example representation for *apple* before (first row) and after (second row) applying the perceptual inference method of Johns and Jones (2012).

sampling procedure described in Andrews et al. (2009).

Inducing feature-topic components from a document collection $\mathcal{D}$ with the extended LDA model just described gives two sets of parameters: word probabilities given components $P_W(w_i|X = x_c)$ for w_i, $i = 1,...,N$, and feature probabilities given components $P_F(f_k|X = x_c)$ for f_k, $k = 1,...,F$. For example, most of the probability mass of component x_{107} would be reserved for the words *apple*, *fruit*, *lemon*, *orange*, *tree* and the features is_red, tastes_sweet, is_round and so on.

Word meaning in this model is represented by the distribution $P_{X|W}$ over the learned components (see Figure 2 for an example). Assuming a uniform distribution over components x_c in $\mathcal{D}$, $P_{X|W}$ can be approximated as:

$$P_{X=x_c|W=w_i} = \frac{P(w_i|x_c)P(x_c)}{P(w_i)} \approx \frac{P(w_i|x_c)}{\sum_{l=1}^{C} P(w_i|x_l)} \quad (4)$$

where C is the total number of components. The model can be also used to infer features for words that were not originally included in the feature norms. The probability distribution $P_{F|W}$ over features given a word w_i is simply inferred by summing over all components x_c for each feature f_k:

$$P_F(f_k|W = w_i) = \sum_{c=1}^{C} P(f_k|x_c)P(x_c|w_i) \quad (5)$$

2.2 Global Similarity Model

Johns and Jones (2012) propose an approach for generating perceptual representations for words by means of global lexical similarity. Their model does not place so much emphasis on the integration of perceptual and linguistic information, rather its main focus is on inducing perceptual representations for words with no perceptual correlates. Their idea is to assume that lexically similar words also share perceptual features and hence it should be possible to transfer perceptual information onto words that have none from their linguistically similar neighbors.

Let $T \in \{1,0\}^{N\times D}$ denote a binary term-document matrix, where each cell records the presence or absence of a term in a document. Let $P \in [0,1]^{N\times F}$ denote a perceptual matrix, representing a probability distribution over features for each word (see Table 1). A word's meaning is represented by the concatenation of its textual and perceptual vectors (see Figure 3). If a word has not been normed, its perceptual vector will be all zeros. Johns and Jones (2012) propose a two-step estimation process for words without perceptual vectors. Initially, a perceptual vector is constructed based on the word's weighted similarity to other words that have non-zero perceptual vectors:

$$\mathbf{p}_{inf} = \sum_{i=1}^{N} \mathbf{t_i} * sim(\mathbf{t_i}, \mathbf{p})^{\lambda} \quad (6)$$

where $\mathbf{p}$ is the representation of a word with a textual vector but an empty perceptual vector, $\mathbf{t}$s are composite representations consisting of textual and perceptual vectors, *sim* is a measure of distributional similarity such as cosine, λ a weighting parameter, and $\mathbf{p}_{inf}$ the resulting inferred representation of the word. The process is repeated a second time, so as to incorporate the inferred perceptual vector in the computation of the inferred vectors of all other words. An example of this inference procedure is illustrated in Figure 3.

$$
\begin{array}{r}
\\
apple
\end{array}
\begin{array}{c}
\begin{array}{cccccccccc}
\ldots & d_{16} & \ldots & d_{322} & \ldots & d_{2470} & \ldots & d_D & \quad a\_fruit & has\_fangs \quad is\_crunchy \ldots is\_yellow \quad is\_red \quad is\_green \quad is\_round \end{array} \\
\begin{bmatrix} \ldots & .006 & \ldots & .003 & \ldots & .1e\text{-}6 & \ldots & 0 \end{bmatrix} \begin{bmatrix} .13 & 0 & .06 & \ldots & .04 & .14 & .09 & .04 \end{bmatrix} \\
\begin{array}{cccccccccc}
\quad k_1 & \quad k_2 & \quad k_3 & \ldots & k_{409} & k_{410} & \quad k_1 & \quad k_2 & \quad k_3 & \ldots \quad k_{409} \quad k_{410} \end{array} \\
\begin{bmatrix} -.003 & -.01 & .002 & \ldots & -.002 & -.01 \end{bmatrix} \begin{bmatrix} .008 & -.03 & -.008 & \ldots & -.02 & -.07 \end{bmatrix}
\end{array}
$$

Figure 4: Example representation for *apple* before (first row) and after (second row) applying CCA.

2.3 Canonical Correlation Analysis

Our third model uses Canonical Correlation Analysis (CCA, Hardoon et al. (2004)) to learn a joint semantic representation from the textual and perceptual views. Given two random variables **x** and **y** (or two sets of vectors), CCA can be seen as determining two sets of basis vectors in such a way, that the correlation between the projections of the variables onto these bases is mutually maximized (Borga, 2001). In effect, the representation-specific details pertaining to the two views of the same phenomenon are discarded and the underlying hidden factors responsible for the correlation are revealed.

In our case the linguistic view is represented by a term-document matrix, $T \in \mathbb{R}^{N \times D}$, containing information about the occurrence of each word in each document. The perceptual view is captured by a perceptual matrix, $P \in [0,1]^{N \times F}$, representing words as a probability distribution over normed features. CCA is concerned with describing linear dependencies between two sets of variables of relatively low dimensionality. Since the correlation between the linguistic and perceptual views may exist in some nonlinear relationship, we used a kernelized version of CCA (Hardoon et al., 2004) which first projects the data into a higher-dimensional feature space and then performs CCA in this new feature space. The two kernel matrices are $K_T = TT'$ and $K_P = PP'$. After applying CCA we obtain two matrices projected onto L basis vectors, $C_t \in \mathbb{R}^{N \times L}$, resulting from the projection of the textual matrix T onto the new basis and $C_p \in \mathbb{R}^{N \times L}$, resulting from the projection of the corresponding perceptual feature matrix. The meaning of a word can thus be represented by its projected textual vector in C_T, its projected perceptual vector in C_P or their concatenation. Figure 4 shows an example of the textual and perceptual vectors for the word *apple* which were used as input for CCA (first row) and their new representation after the projection onto new basis vectors (second row).

The CCA model as sketched above will only obtain full representations for words with perceptual features available. One solution would be to apply the method from Johns and Jones (2012) to infer the perceptual vectors and then perform CCA on the inferred vectors. Another approach which we assess experimentally (see Section 4) is to create a perceptual vector for a word that has none from its k-most (textually) similar neighbors, simply by taking the average of their perceptual vectors. This inference procedure can be applied to the original vectors or the projected vectors in C_T and C_P, respectively, once CCA has taken place.

2.4 Discussion

Johns and Jones (2012) primarily present a model of perceptual inference, where textual data is used to infer perceptual information for words not included in feature norms. There is no means in this model to obtain a joint representation resulting from the mutual influence of the perceptual and textual views. As shown in the example in Figure 3 the textual vector on the left-hand side does not undergo any transformation whatsoever. The generative model put forward by Andrews et al. (2009) learns meaning representations by simultaneously considering documents and features. Rather than simply adding perceptual information to textual data it integrates both modalities jointly in a *single* representation which is desirable, at least from a cognitive perspective. It is unlikely that we have separate representations for different aspects of word meaning (Rogers et al., 2004). Similarly to Johns and Jones (2012), Andrews et al's (2009) feature-topic model can also infer perceptual representations for words that have none. The inference is performed automatically in an implicit manner during component induction.

In CCA, textual and perceptual data represent two different views of the same objects and the model operates on these views *directly* without combining or manipulating any of them a priori. Instead, the combination of the two modalities is realized via

correlating the linear relationships between them. A drawback of the model lies in the need of additional methods for inferring perceptual representations for words not available in feature norms.

3 Experimental Setup

Data All our experiments used a lemmatized version of the British National Corpus (BNC) as a source of textual information. The feature norms of McRae et al. (2005) were used as a proxy for perceptual information. The BNC comprises 4,049 texts totalling approximately 100 million words. McRae et al.'s feature norms consist of 541 words and 2,526 features; 824 of these features occur with at least two different words.

Evaluation Tasks Our evaluation experiments compared the models discussed above on three tasks. Two of them 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., 1998).[1] 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 cue 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 models (see the following section for details on the similarity measures we employed). The norms contain 63,619 unique normed cue-associate pairs in total. Of these, 25,968 pairs were covered by all models and 520 appeared in McRae et al.'s (2005) norms. Using correlation analysis, we examined the degree of linear relationship between the human cue-associate probabilities and the automatically derived similarity values.

Our word similarity experiments used the WordSimilarity-353 test collection (Finkelstein et al., 2002)[2] which consists of relatedness judgments for word pairs. For each pair, a similarity judgment (on a scale of 0 to 10) was elicited from 13 or 16 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 to generate the first word that came to mind in response to a cue word. The collection contains similarity ratings for 353 word pairs. Of these, 76 pairs appeared in our corpus and 3 in McRae et al.'s (2005) norms. Again, we evaluated how well model produced similarities correlate with human ratings. Throughout this paper we report correlation coefficients using Pearson's r.

Our third task assessed the models' ability to infer perceptual vectors for words that have none. To do this, we conducted 10-fold cross-validation on McRae et al.'s (2005) norms. We treated the perceptual vectors in each test fold as unseen, and used the data in the corresponding training fold together with the models presented in Section 2 to infer them. Then, for each word, we examined how close the inferred vector was to the actual one, via correlation analysis.

Model Parameters The feature-topic model has a few parameters that must be instantiated. These include, C, the number of predefined components and the priors α, β, and γ. Following Andrews et al. (2009), the components C were set to 350.[3] A vague inverse gamma prior was placed on α, β, and γ.[4] To measure word similarity within this model, we adopt Griffiths et al.'s (2007) definition. 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 $P(w_2|w_1)$. Assuming that both w_1 and w_2 came from a single component, $P(w_2|w_1)$ can be estimated as:

$$P(w_2|w_1) = \sum_{c=1}^{C} P(w_2|x_c)P(x_c|w_1)$$
$$P(x_c|w_1) \propto P(w_1|x_c)P(x_c) \quad (7)$$

[1] Available at http://www.usf.edu/Freeassociation.

[2] Available at http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/.

[3] As we explain in Section 4 the feature-topic model was compared to a vanilla LDA model trained on the BNC only. For that model, C was set to 250.

[4] That is $P(\bullet) = exp(-\frac{1}{\bullet})\bullet^{-2}$.

where $P(x_c)$ is uniform, a single component x_c is sampled from the distribution $P(x_c|w_1)$, and an overall estimate is obtained by averaging over all C components.

Johns and Jones' (2012) model uses binary textual vectors to represent word meaning. If the word is present in a given document, that vector element is coded as one; if it is absent, it is coded as zero. We built a binary term-document matrix from the BNC over 14,000 lemmas. The value of the similarity weighting parameter λ was set to the same values reported by Johns and Jones (λ_1=3 for Step 1 and $\lambda_2 = 13$ for Step 2).

For the CCA model, we represented the textual view with a term-document co-occurrence matrix. Matrix cells were set to their tf-idf values.[5] The textual and perceptual matrices were projected onto 410 vectors. As mentioned in Section 2.3, CCA does not naturally lend itself to inferring perceptual vectors, yet a perceptual vector for a word can be created from its k-nearest neighbors. We inferred a perceptual vector by averaging over the perceptual vectors of the word's k most similar words; textual similarity between two words was measured using the cosine of the angle of the two vectors representing them. To find the optimal value for k, we used one third of Nelson's (1998) cues as development set. The highest correlation was achieved with $k = 2$ when the perceptual vectors were created prior to CCA and $k = 8$ when they were inferred on the projected textual and perceptual matrices.

4 Results

Our experiments were designed to answer three questions: (1) Does the integration of perceptual and textual information yield a better fit with behavioral data compared to a model that considers only one data source? (2) What is the best way to integrate the two modalities, e.g., via simple concatenation or jointly? (3) How accurately can we approximate the perceptual information when the latter is absent?

To answer the first question, we assessed the models' performance when textual and perceptual information are both available. The results in Table 2 are thus computed on the subset of Nelson's (1998) norms (520 cue-associate pairs) that also appeared in McRae et al. (2005) and for which a perceptual vector was present. The table shows different instantiations of the three models depending on the type of modality taken into account: textual, perceptual or both.

Models	Modality	Pearson's r
Feature-topic	+t +p	.35
Feature-topic	+t −p	.12
Feature-topic	−t +p	.22
Global similarity	+t +p	.23
Global similarity	+t −p	.11
Global similarity	−t +p	.22
CCA	+t +p	.32
CCA	+t −p	.14
CCA	−t +p	.29
Upper Bound	—	.91

Table 2: Performance of feature-topic, global similarity, and CCA models on a subset of the Nelson et al. (1998) norms when taking into account the textual and perceptual modalities on their own (+t−p and −t+p) and in combination (+t+p). All correlation coefficients are statistically significant ($p < 0.01$).

As can be seen, Andrews et al.'s (2009) feature-topic model provides a better fit with the association data when both modalities are taken into account (+t+p). A vanilla LDA model constructed solely on the BNC (+t−p) or McRae et al.'s (2005) feature norms (−t+p) yields substantially lower correlations. We observe a similar pattern with Johns and Jones' (2012) global similarity model. Concatenation of perceptual and textual vectors yields the best fit with the norming data, relying on perceptual information alone (−t+p) comes close, whereas textual information on its own seems to have a weaker effect (+t−p).[6] The CCA model takes perceptual and textual information as input in order to find a projection onto basis vectors that are maximally correlated. Although by definition the CCA model must operate on the two views, we can nevertheless isolate the contribution of each modality by considering the vectors resulting from the projection of the tex-

[5] Experiments with a binarized version of the term-document matrix consistently performed worse.

[6] In this evaluation setting, the model does not infer any perceptual representations; perceptual vectors are taken directly from McRae et al. (2005).

tual matrix (+t−p), the perceptual matrix (−t+p) or their concatenation (+t+p). We obtain best results with the latter representation; again we observe that the perceptual information is more dominant.

Overall we find that the feature-topic model and CCA perform best. In fact the correlations achieved by the two models do not differ significantly, using a t-test (Cohen and Cohen, 1983). The performance of the global similarity model is significantly worse than the feature-topic model and CCA ($p < 0.01$). Recall that the feature-topic model (+t+p) represents words as distributions over components, whereas the global similarity model simply concatenates the textual and perceptual vectors. The same input is also given to CCA which in turn attempts to interpret the data by inferring common relationships between the two views. In sum, we can conclude that the higher correlation with human judgments indicates that integrating textual and perceptual modalities jointly is preferable to concatenation.

However, note that all models in Table 2 fall short of the human upper bound which we measured by calculating the *reliability* of Nelson et al.'s (1998) norms. Reliability estimates the likelihood of a similarly-composed group of participants presented with the same task under the same circumstances producing identical results. We split the collected cue-associate pairs randomly into two halves and computed the correlation between them; this correlation was averaged across 200 random splits. These correlations were adjusted by applying the Spearman-Brown prediction formula (Voorspoels et al., 2008).

The results in Table 2 are computed on a small fraction of Nelson et al.'s (1998) norms. One might even argue that the comparison is slightly unfair as the global similarity model is more geared towards inferring perceptual vectors rather than integrating the two modalities in the best possible way. To gain a better understanding of the models' behavior and to allow comparisons on a larger dataset and more equal footing, we also report results on the entire dataset (20,556 cue-associate pairs).[7] This entails that the models will infer perceptual vectors for the words that are not attested in McRae et al.'s norms. Recall from Section 2.3 that CCA does not have a dedicated inference mechanism. We thus experimented with three options (a) interfacing the inference method of Johns and Jones (2012) with CCA (global similarity $\ll$ CCA) (b) creating a perceptual vector from the words' k-nearest neighbors before (k-NN $\ll$ CCA) or (c) after CCA takes place (CCA $\ll$ k-NN).

Models	Pearson's r
Feature-topic	.15
Global similarity	.03
Global similarity $\ll$ CCA	.12
k-NN $\ll$ CCA	.11
CCA $\ll$ k-NN	.12
Upper Bound	.96

Table 3: Performance of the feature-topic, global similarityand CCA models on the Nelson et al. (1998) norms (entire dataset). All correlation coefficients are statistically significant ($p < 0.01$).

Our results are summarized in Table 3. The upper bound was estimated in the same fashion as for the smaller dataset. Despite being statistically significant ($p < 0.01$), the correlation coefficients are lower. This is hardly surprising as perceptual information is approximate and in several cases likely to be wrong. Interestingly, we observe similar modeling trends, irrespective of whether the models are performing perceptual inference or not. The feature-topic model achieves the best fit with the data, followed by CCA. The inference method here does not seem to have much of an impact: CCA $\ll$ k-NN does as well as global similarity $\ll$ CCA. This is perhaps expected as the inference procedure adopted by Johns and Jones (2012) is a generalization of our k-nearest neighbor approach. The global similarity model performs worst; we conjecture that this is due to the way semantic information is integrated rather than the inference method itself. CCA works with similar input, yet achieves better correlations with the human data, due to its ability to represent the commonalities shared by the two modalities. Taken together the results in Tables 2 and 3 provide an answer to our second question. Models that capture latent information shared between the two modalities

[7] This excludes the data used as development set for tuning the k-nearest neighbors for CCA.

Models	Pearson's r
Feature-topic	.17
Global similarity	.25
Global similarity $\ll$ CCA	.21
k-NN $\ll$ CCA	.19
CCA $\ll$ k-NN	.13

Table 4: Mean correlation coefficients between original and inferred feature vectors in McRae et al.'s (2005) norms.

Models	Pearson's r
Feature-topic	.35
Global similarity	.08
Global similarity $\ll$ CCA	.38
k-NN $\ll$ CCA	.39
CCA $\ll$ k-NN	.28
Upper Bound	.98

Table 5: Model performance on predicting word similarity. All correlation coefficients are statistically significant ($p < 0.01$), except for the global similarity model.

create more accurate semantic representations compared to simply treating the two as independent data sources.

In order to isolate the influence of the inference method from the resulting semantic representation we evaluated the inferred perceptual vectors on their own by computing their correlation with the original feature distributions in McRae et al.'s (2005) norms. The correlation coefficients are reported in Table 4 and were computed by averaging the coefficients obtained for individual words. Here, the global similarity model achieves the highest correlation, and for a good reason. It is the only model with an emphasis on inference, the other two models do not have such a dedicated mechanism. CCA has in fact none, whereas in the feature-topic model the inference of missing perceptual information is a by-product of the generative process. The results in Table 4 indicate that the perceptual vectors are not reconstructed very accurately (the highest correlation coefficient is .25) and that better inference mechanisms are required for perceptual information to have a positive impact on semantic representation.

In Table 5 we examine the models' performance on semantic similarity rather than association using the WordSimilarity-353 dataset (Finkelstein et al., 2002). The models were evaluated on 76 word pairs that appeared in the BNC. We inferred the perceptual vectors for 51 words. We computed the upper bound using the reliability method described earlier. Again, the joint models achieve better results than the simple concatenation model. The feature-topic and CCA models perform comparably, with the global similarity model lagging substantially behind. In sum, our results indicate that the issue of how to best integrate the two modalities has a greater impact on the resulting semantic representations compared to the mechanism by which missing perceptual information is inferred.

5 Conclusions

In this paper, we have presented a comparative study of semantic representation models which compute word meaning on the basis of linguistic and perceptual information. The models differ in terms of the mechanisms by which they integrate the two modalities. In the feature-topic model (Andrews et al., 2009), the textual and perceptual views are integrated via a set of latent components that are inferred from the *joint* distribution of textual words and perceptual features. The model based on Canonical Correlation Analysis (Hardoon et al., 2004) integrates the two views by deriving a *consensus* representation based on the correlation between the linguistic and perceptual modalities. Johns and Jones' (2012) similarity-based model simply concatentates the two representations. In addition, it uses the linguistic representations of words to infer perceptual information when the latter is absent.

Experiments on word association and similarity show that all models benefit from the integration of perceptual data. We also find that joint models are superior as they obtain a closer fit with human judgments compared with an approach that simply concatenates the two views. We have also examined how these models perform on the perceptual inference task which has implications for the wider applicability of grounded semantic representation models. Johns and Jones' (2012) inference mechanism goes some way towards reconstructing the information contained in the feature norms, however, further

work is needed to achieve representations accurate enough to be useful in semantic tasks.

In this paper we have used McRae et al.'s (2005) norms without any extensive feature engineering other than applying a frequency cut-off. In the future we plan to experiment with feature selection methods in an attempt to represent perceptual information more succinctly. For example, it may be that different features are appropriate for different word classes (e.g., color versus event denoting nouns). Although feature norms are a useful first approximation of perceptual data, the effort involved in eliciting them limits the scope of any computational model based on normed data. A natural avenue for future work would be to develop semantic representation models that exploit perceptual data that is both naturally occurring and easily accessible (e.g., images, physical simulations).

Acknowledgments We are grateful to Brendan Johns for his help with the re-implementation of his model. Thanks to Frank Keller and Michael Roth for their input on earlier versions of this work, Ioannis Konstas for his help with the final version, and members of the ILCC at the School of Informatics for valuable discussions and comments. We acknowledge the support of EPSRC through project grant EP/I032916/1.

References

M. Andrews, G. Vigliocco, and D. Vinson. 2009. Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116(3):463–498.

Lawrence Barsalou. 1999. Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22:577–609.

Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology*, 59:617–845.

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

Magnus Borga. 2001. Canonical Correlation - a Tutorial, January.

M. H. Bornstein, 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.

Elia Bruni, Giang Binh Tran, and Marco Baroni. 2011. Distributional Semantics from Text and Images. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 22–32, Edinburgh, UK, July. Association for Computational Linguistics.

Freddy Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of the 6th EMNLP*, pages 109–117, Seattle, WA.

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

Yansong Feng and Mirella Lapata. 2010. Visual Information in Semantic Representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, Los Angeles, California, June. Association for Computational Linguistics.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1):116–131, January.

Arthur M. Glenberg and Michael P. Kaschak. 2002. Grounding Language in Action. *Psychonomic Bulletin and Review*, 9(3):558–565.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

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

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. 2004. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664.

H Hotelling. 1936. Relations between Two Sets of Variates. *Biometrika*, 28:312–377.

Steve R. Howell, Damian Jankowicz, and Suzanna Becker. 2005. A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning. *Journal of Memory and Language, 53(2), 258-276*, 53(2):258–276.

Brendan T. Johns and Michael N. Jones. 2012. Perceptual Inference through Global Lexical Similarity. *Topics in Cognitive Science*, 4(1):103–120.

B. Landau, L. Smith, and S. Jones. 1998. Object Perception and Object Naming in Early Development. *Trends in Cognitive Science*, 27:19–24.

T. Landauer 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.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 768–774, Montréal, Canada.

K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. 2005. Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37(4):547–559, November.

D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. 1998. The University of South Florida Word Association, Rhyme, and Word Fragment Norms.

C. Perfetti. 1998. The Limits of Co-occurrence: Tools and Theories in Language Research. *Discourse Processes*, (25):363–377.

Terry Regier. 1996. *The Human Semantic Potential*. MIT Press, Cambridge, MA.

T. T. Rogers, 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.

G Salton, A Wang, and C Yang. 1975. A Vector-space Model for Information Retrieval. *Journal of the American Society for Information Science*, 18:613–620.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–124.

Mark Steyvers. 2010. Combining Feature Norms and Text Data with Topic Models. *Acta Psychologica*, 133(3):234–342.

Wouter Voorspoels, Wolf Vanpaemel, and Gert Storms. 2008. Exemplars and Prototypes in Natural Language Concepts: A Typicality-based Evaluation. *Psychonomic Bulletin & Review*, 15:630–637.