Multi-Modal Models for Concrete and Abstract Concept Meaning

Felix Hill
Computer Laboratory
University of Cambridge
fh295@cam.ac.uk

Roi Reichart
Technion - IIT
Haifa, Israel
roiri@ie.technion.ac.il

Anna Korhonen
Computer Laboratory
University of Cambridge
alk23@cam.ac.uk

Abstract

Multi-modal models that learn semantic representations from both linguistic and perceptual input outperform language-only models on a range of evaluations, and better reflect human concept acquisition. Most perceptual input to such models corresponds to concrete noun concepts and the superiority of the multi-modal approach has only been established when evaluating on such concepts. We therefore investigate which concepts can be effectively learned by multi-modal models. We show that concreteness determines both which linguistic features are most informative and the impact of perceptual input in such models. We then introduce *ridge regression* as a means of propagating perceptual information from concrete nouns to more abstract concepts that is more robust than previous approaches. Finally, we present *weighted gram matrix combination*, a means of combining representations from distinct modalities that outperforms alternatives when both modalities are sufficiently rich.

1 Introduction

What information is needed to learn the meaning of a word? Children learning words are exposed to a diverse mix of information sources. These include clues in the language itself, such as nearby words or speaker intention, but also what the child perceives about the world around it when the word is heard. Learning the meaning of words requires not only a sensitivity to both linguistic and perceptual input, but also the ability to process and combine information from these modalities in a productive way.

Many computational semantic models represent words as real-valued vectors, encoding their relative frequency of occurrence in particular forms and contexts in linguistic corpora (Sahlgren, 2006; Turney et al., 2010). Motivated both by parallels with human language acquisition and by evidence that many word meanings are *grounded* in the perceptual system (Barsalou et al., 2003), recent research has explored the integration into text-based models of input that approximates the visual or other sensory modalities (Silberer and Lapata, 2012; Bruni et al., 2014). Such models can learn higher-quality semantic representations than conventional corpus-only models, as evidenced by a range of evaluations.

However, the majority of perceptual input for the models in these studies corresponds directly to concrete noun concepts, such as *chocolate* or *cheeseburger*, and the superiority of the multi-modal over the corpus-only approach has only been established when evaluations include such concepts (Leong and Mihalcea, 2011; Bruni et al., 2012; Roller and Schulte im Walde, 2013; Silberer and Lapata, 2012). It is thus unclear if the multi-modal approach is effective for more abstract words, such as *guilt* or *obesity*. Indeed, since empirical evidence indicates differences in the representational frameworks of both concrete and abstract concepts (Paivio, 1991; Hill et al., 2013), and verb and noun concepts (Markman and Wisniewski, 1997), perceptual information may not fulfill the same role in the representation of the various concept types. This potential challenge to the multi-modal approach is of particular practical importance since concrete nouns constitute only a small proportion of the open-class, meaning-bearing

Transactions of the Association for Computational Linguistics, 2 (2014) 285–296. Action Editor: Rada Mihalcea.
Submitted 12/2013; Revised 6/2014; Published 10/2014.

words in everyday language (Section 2).

In light of these considerations, this paper addresses three questions: (1) Which information sources (modalities) are important for acquiring concepts of different types? (2) Can perceptual input be propagated effectively from concrete to more abstract words? (3) What is the best way to combine information from the different sources?

We construct models that acquire semantic representations for four sets of concepts: concrete nouns, abstract nouns, concrete verbs and abstract verbs. The linguistic input to the models comes from the recently released Google Syntactic N-Grams Corpus (Goldberg and Orwant, 2013), from which a selection of linguistic features are extracted. Perceptual input is approximated by data from the McRae et al. (2005) norms, which encode perceptual properties of concrete nouns, and the ESPGame dataset (Von Ahn and Dabbish, 2004), which contains manually generated descriptions of 100,000 images.

To address (1) we extract representations for each concept type from combinations of information sources. We first focus on different classes of linguistic features, before extending our models to the multi-modal context. While linguistic information overall effectively reflects the meaning of all concept types, we show that features encoding syntactic patterns are only valuable for the acquisition of abstract concepts. On the other hand, perceptual information, whether directly encoded or propagated through the model, plays a more important role in the representation of concrete concepts.

In addressing (2), we propose *ridge regression* (Myers, 1990) as a means of propagating features from concrete nouns to more abstract concepts. The regularization term in ridge regression encourages solutions that generalize well across concept types. We show that ridge regression effectively propagates perceptual information to abstract nouns and concrete verbs, and is overall preferable to both linear regression and the method of Johns and Jones (2012) applied to a similar task by Silberer and Lapata (2012). However, for all propagation methods, the impact of integrating perceptual information depends on the concreteness of the target concepts. Indeed, for abstract verbs, the most abstract concept type in our evaluations, perceptual input actually degrades representation quality. This highlights the need to consider the concreteness of the target domain when constructing multi-modal models.

To address (3), we present various means of combining information from different modalities. We propose *weighted gram matrix combination*, a technique in which representations of distinct modalities are mapped to a space of common dimension where coordinates reflect proximity to other concepts. This transformation, which has been shown to enhance semantic representations in the context of verb-clustering (Reichart and Korhonen, 2013), reduces representation sparsity and facilitates a product-based combination that results in greater inter-modal dependency. Weighted gram matrix combination outperforms alternatives such as concatenation and Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) when combining representations from two similarly rich information sources.

In Section 3, we present experiments with linguistic features designed to address question (1). These analyses are extended to multi-modal models in Section 4, where we also address (2) and (3). We first discuss the relevance of concreteness and part-of-speech (lexical function) to concept representation.

2 Concreteness and Word Meaning

A large and growing body of psychological evidence indicates differences between abstract and concrete concepts.[1] It has been shown that concrete words are more easily learned, remembered and processed than abstract words (Paivio, 1991; Schwanenflugel and Shoben, 1983), while neuroimaging studies demonstrate differences in brain activity when subjects are presented with stimuli corresponding to the two concept types (Binder et al., 2005).

The abstract/concrete distinction is important to computational semantics for various reasons. While many models construct representations of concrete words (Andrews et al., 2009; Landauer and Dumais, 1997), abstract words are in fact far more common in everyday language. For instance, based on an analysis of those noun concepts in the University of South Florida dataset (USF) and their occurrence in the British National Corpus (BNC) (Leech et al., 1994), 72% of noun tokens in corpora are rated by human

[1] Here *concreteness* is understood intuitively, as per the psychological literature (Rosen, 2001; Gallese and Lakoff, 2005).

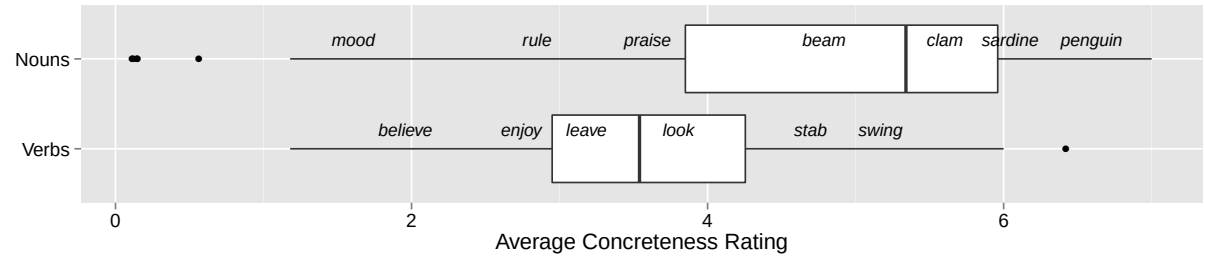

Figure 1: Boxplot of concreteness distributions for noun and verb concepts in the USF data, with selected example concepts. The bold vertical line is the mean, boxes extend from the first to the third quartile, and dots represent outliers.

judges as more abstract than the noun *war*, a concept that many would already consider quite abstract.[2]

The recent interest in multi-modal semantics further motivates a principled modelling approach to lexical concreteness. Many multi-modal models implicitly distinguish concrete and abstract concepts since their perceptual input corresponds only to concrete words (Bruni et al., 2012; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013). However, given that many abstract concepts express relations or modifications of concrete concepts (Gentner and Markman, 1997), it is reasonable to expect that perceptual information about concrete concepts could also enhance the quality of more abstract representations in an appropriately constructed model.

Moreover, concreteness is closely related to more functional lexical distinctions, such as those between adjectives, nouns and verbs. An analysis of the USF dataset, which includes concreteness ratings for over 4,000 words collected from thousands of participants, indicates that on average verbs (mean concreteness, 3.64) are considered more abstract than nouns (mean concreteness, 4.91), an effect illustrated in Figure 1. This connection between lexical function and concreteness suggests that a sensitivity to concreteness could improve models that already make principled distinctions between words based on their part-of-speech (POS) (Im Walde, 2006; Baroni and Zamparelli, 2010).

Although the focus of this paper is on multi-modal models, few conventional semantic models make principled distinctions between concepts based on function or concreteness. Before turning to the multi-modal case, we thus investigate whether these distinctions are pertinent to text-only models.

[2]This sample covers 15.2% of all noun tokens in the BNC.

3 Concreteness and Linguistic Features

It has long been known that aspects of word meaning can be inferred from nearby words in corpora. Approaches that exploit this fact are often called *distributional models* (Sahlgren, 2006; Turney et al., 2010). We take a distributional approach to learning linguistic representations. The advantage of using distributional methods to learn representations from corpora versus approaches that rely on knowledge bases (Pedersen et al., 2004; Leong and Mihalcea, 2011) is that they are more scalable, easily applicable across languages and plausibly reflect the process of human word learning (Landauer and Dumais, 1997; Griffiths et al., 2007). We group distributional features into three classes to test which forms of linguistic information are most pertinent to the abstract/concrete and verb/noun distinctions.

All features are extracted from The Google Syntactic N-grams Corpus. The dataset contains counted dependency-tree fragments for over 10bn words of the English Google Books Corpus.

3.1 Feature Classes

Lexical Features Our lexical features are the co-occurrence counts of a concept word with each of the other 2,529 concepts in the USF data. Co-occurrences are counted in a 5-word window, and, as elsewhere (Erk and Padó, 2008), weighted by pointwise mutual information (PMI) to control for the underlying frequency of both concept and word.

POS-tag Features Many words function as more than one POS, and this variation can be indicative of meaning (Manning, 2011). For example, deverbal

	Context	Example
Noun Concepts	indirect object	*gave it to the **man***
	direct object	*gave the **pie** to him*
	subject	*the **man** grinned*
	in PP	*was in his **mouth***
	adject. modifier	*the portly **man***
Verb Concepts	infinitive clause	*to **eat** is human*
	transitive	*he **bit** the steak*
	intransitive	*he **salivated***
	distransitive	***put** jam on the toast*
	phrasal verb	*he **gobbled** it up*
	infinitival comp.	*he **wants** to snooze*
	clausal comp.	*I **bet** he won't diet*

Table 1: Grammatical features for noun/verb concepts

nouns, such as *shiver* or *walk*, often refer to processes rather than entities. To capture such effects, we count the frequency of occurrence with the POS categories *ajdective, adverb, noun* and *verb*.

Grammatical Features Grammatical role is a strong predictor of semantics (Gildea and Jurafsky, 2002). For instance, the subject of transitive verbs is more likely to refer to an animate entity than a noun chosen at random. Syntactic context also predicts verb semantics (Kipper et al., 2008). We thus count the frequency of nouns in a range of (non-lexicalized) syntactic contexts, and of verbs in one of the six most common *subcategorization-frame* classes as defined in Van de Cruys et al. (2012). These contexts are detailed in Table 1.

3.2 Evaluation Sets

We create evaluation sets of abstract and concrete concepts, and introduce a complementary dichotomy between nouns and verbs, the two POS categories most fundamental to propositional meaning. To construct these sets, we extract nouns and verbs from word pairs in the USF data based on their majority POS-tag in the lemmatized BNC (Leech et al., 1994), excluding any word not assigned to either of the POS categories in more than 70% of instances. From the resulting 2175 nouns and 354 verbs, the abstract-concrete distinction is drawn by ordering words according to concreteness and sampling at random from the first and fourth quartiles. Any concrete nouns not occurring in the McRae et al. (2005) Property Norm dataset were also excluded.

Concept Type	Words	Pairs	Examples
concrete nouns	303	1280	*yacht, cup*
abstract nouns	100	295	*fear, respect*
all nouns	403	1716	*cup, respect*
concrete verbs	50	66	*kiss, launch*
abstract verbs	50	127	*differ, obey*
all verbs	100	221	*kiss, differ*

Table 2: Evaluation sets used throughout. *All nouns* and *all verbs* are the union of abstract and concrete subsets and mixed *abstract-concrete* or *concrete-abstract* pairs.

For each list of concepts L = *concrete nouns, concrete verbs, abstract nouns, abstract verbs*, together with lists *all nouns* and *all verbs*, a corresponding set of pairs $\{(w_1, w_2) \in USF : w_1, w_2 \in L\}$ is defined for evaluation. These details are summarized in Table 2. Evaluation lists, sets of pairs and USF scores are downloadable from our website.

3.3 Evaluation Methodology

All models are evaluated by measuring correlations with the *free-association* scores in the USF dataset (Nelson et al., 2004). This dataset contains the free-association strength of over 150,000 word pairs.[3] These data reflect the cognitive proximity of concepts and have been widely used in NLP as a gold-standard for computational models (Andrews et al., 2009; Feng and Lapata, 2010; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013).

For evaluation pairs (c_1, c_2) we calculate the cosine similarity between our learned feature representations for c_1 and c_2, a standard measure of the proximity of two vectors (Turney et al., 2010), and follow previous studies (Leong and Mihalcea, 2011; Huang et al., 2012) in using Spearman's ρ as a measure of correlation between these values and our gold-standard.[4] All representations in this section are combined by concatenation, since the present focus is not on combination methods.[5]

[3]Free-association strength is measured by presenting subjects with a cue word and asking them to produce the *first word they can think of that is associated with that cue word.*

[4]We consider Spearman's ρ, a non-parametric ranking correlation, to be more appropriate than Pearson's r for free association data, which is naturally skewed and non-continuous.

[5]When combining multiple representations we normalize

Feature Type	All Nouns	Conc. Nouns	Abs. Nouns	All Verbs	Conc. Verbs	Abs. Verbs
(1) Lexical	0.168*	**0.199***	**0.248***	**0.173***	**0.268***	0.109
(2) POS-tag	0.059*	0.012	0.119*	0.052	-0.074	**0.123**
(3) Grammatical	0.078*	0.027	0.121*	0.009	-0.017	0.114
(1)+(2)+(3)	**0.182 ***	0.181*	0.247*	0.172*	0.267*	0.108

Table 3: Spearman correlation ρ of cosine similarity between vector representations derived from three feature classes with USF scores. * indicates statistically significant correlations ($p < 0.05$).

3.4 Results

The performance of each feature class on the evaluation sets is detailed in Table 3. When all linguistic features are included, performance is somewhat better on noun concepts ($\rho = 0.182$) than verbs ($\rho = 0.172$). However, while correlations are significant on concrete ($\rho = 0.181$) and abstract nouns ($\rho = 0.247$) and concrete verbs, the effect is not significant on abstract verbs (although it is on verbs overall). The highest correlations for the linguistic features together are on abstract nouns ($\rho = 0.247$) and concrete verbs ($\rho = 0.267$). Referring back to the continuum in Figure 1, it is possible that there is an optimum concreteness level, exhibited by abstract nouns and concrete verbs, at which conceptual meaning is best captured by linguistic models.

The results indicate that the three feature classes convey distinct information. It is perhaps unsurprising that lexical features produce the best performance in the majority of cases; the value of lexical co-occurrence statistics in conveying word meaning is expressed in the well known *distributional hypothesis* (Harris, 1954). More interestingly, on abstract concepts the contribution of POS-tag (nouns, $\rho = 0.119$; verbs, $\rho = 0.123$) and grammatical features (nouns, $\rho = 0.121$; verbs, $\rho = 0.114$) is notably higher than on the corresponding concrete concepts. The importance of such features to modelling free-association between abstract concepts suggests that they may convey information about how concepts are (subjectively) organized and interrelated in the minds of language users, independent of their realisation in the physical world. Indeed, since abstract representations rely to a lesser extent than concrete representations on perceptual input (Section 4), it is perhaps unsurprising that more of their meaning is reflected in subtle linguistic patterns.

The results in this section demonstrate that different information is required to learn representations for abstract and concrete concepts and for noun and verb concepts. In the next section, we investigate how perceptual information fits into this equation.

each representation, then concatenate and then renormalize.

4 Acquiring Multi-Modal Representations

As noted in Section 2, there is experimental evidence that perceptual information plays a distinct role in the representation of different concept types. We explore whether this finding extends to computational models by integrating such information into our corpus-based approaches. We focus on two aspects of the integration process. *Propagation*: Can models infer useful information about abstract nouns and verbs from perceptual information corresponding to concrete nouns? And *combination*: How can linguistic and (propagated or actual) perceptual information be integrated into a single, multi-modal representation? We begin by introducing the two sources of perceptual information.

4.1 Perceptual Information Sources

The McRae Dataset The McRae et al. (2005) Property Norms dataset is commonly used as a perceptual information source in cognitively-motivated semantic models (Kelly et al., 2010; Roller and Schulte im Walde, 2013). The dataset contains properties of over 500 concrete noun concepts produced by 30 human annotators. The proportion of subjects producing each property gives a measure of the strength of that property for a given concept. We encode this data in vectors with coordinates for each of the 2,526 properties in the dataset. A concept representation contains (real-valued) feature strengths in places corresponding to the features of that concept and zeros elsewhere. Having defined the concrete noun evaluation set as the 303 concepts found in both the USF and McRae datasets, this information is available for all concrete nouns.

The ESP-Game Dataset To complement the cognitively-driven McRae data with a more explicitly visual information source, we also extract information from the ESP-Game dataset (Von Ahn and Dabbish, 2004) of 100,000 photographs, each annotated with a list of entities depicted in that image. This input enables connections to be made between concepts that co-occur in scenes, and thus might be experienced together by language learners at a given time. Because we want our models to reflect human concept learning in inferring conceptual knowledge from comparatively unstructured data, we use the ESP-Game dataset in preference to resources such as ImageNet (Deng et al., 2009), in which the conceptual hierarchy is directly encoded by expert annotators. An additional motivation is that ESP-Game was produced by crowdsourcing a simple task with untrained annotators, and thus represents a more scalable class of data source.

We represent the ESP-Game data in 100,000 dimensional vectors, with co-ordinates corresponding to each image in the dataset. A concept representation contains a 1 in any place that corresponds to an image in which the concept appears, and a 0 otherwise. Although it is possible to portray actions and processes in static images, and several of the ESP-Game images are annotated with verb concepts, for a cleaner analysis of the information propagation process we only include ESP input in our models for the concrete nouns in the evaluation set.

The data encoding outlined above results in perceptual representations of dimension $\approx 100,000$, for which, on average, fewer than 0.5% of entries are non-zero [6]. In contrast, in our full linguistic representations of nouns (dimension $\approx 4,000$) and verbs (dimension $\approx 8,000$) (Section 3), an average of 24% of entries are non-zero. One of the challenges for the propagation and combination methods described in the following subsections is therefore to manage the differences in dimension and sparsity between linguistic and perceptual representations.

4.2 Information Propagation

Johns and Jones Silberer and Lapata (2012) apply a method designed by Johns and Jones (2012) to infer quasi-perceptual representations for a concept in the case that actual perceptual information is not available. Translating their approach to the present context, for verbs and abstract nouns we infer quasi-perceptual representations based on the perceptual features of concrete nouns that are nearby in the semantic space defined by the linguistic features.

In the first step of their two-step method, for each abstract noun or verb $\mathbf{k}$, a quasi-perceptual representation is computed as an average of the perceptual representations of the concrete nouns, weighted by the proximity between these nouns and $\mathbf{k}$

$$\mathbf{k}^p = \sum_{\mathbf{c} \in \bar{C}} S(\mathbf{k}^l, \mathbf{c}^l)^\lambda \cdot \mathbf{c}^p$$

where $\bar{C}$ is the set of concrete nouns, $\mathbf{c}^p$ and $\mathbf{k}^p$ are the perceptual representations for $\mathbf{c}$ and $\mathbf{k}$ respectively, and $\mathbf{c}^l$ and $\mathbf{k}^l$ the linguistic representations. The exponent parameter λ reflects the learning rate.

Following Johns and Jones (2012), we define the proximity function S between noun concepts to be cosine similarity. However, because our verb and noun representations are of different dimension, we take verb-noun proximity to be the PMI between the two words in the corpus, with co-occurrences counted within a 5-word window.

In step two, the initial quasi-perceptual representations are inferred for a second time, but with the weighted average calculated over the perceptual or initial quasi-perceptual representations of all other words, not just concrete nouns. As with Johns and Jones (2012), we set the learning rate parameter λ to be 3 in the first step and 13 in the second.

Ridge Regression As an alternative propagation method we propose ridge regression (Myers, 1990). Ridge regression is a variant of least squares regression in which a regularization term is added to the training objective to favor solutions with certain properties. Here we apply it to learn parameters for linear maps from linguistic representations of concrete nouns to features in their perceptual representations. For concepts with perceptual representations of dimension n_p, we learn n_p linear functions $f_i : \mathbb{R}^{n_l} \rightarrow \mathbb{R}$ that map the linguistic representations (of dimension n_l) to a particular perceptual feature i. These functions are then applied together to map

[6] The ESP-Game and McRae representations are of approximately equal sparsity.

the linguistic representations of abstract nouns and verbs to full quasi-perceptual representations.[7]

As our model is trained on concrete nouns but applied to other concept types, we do not wish the mapping to reflect the training data too faithfully. To mitigate against this we define our regularization term as the Euclidian l_2 norm of the inferred parameter vector. This term ensures that the regression favors lower coefficients and a smoother solution function, which should provide better generalization performance than simple linear regression. The objective for learning the f_i is then to minimize

$$\|\mathbf{a}X - Y_i\|_2^2 + \|\mathbf{a}\|_2^2$$

where $\mathbf{a}$ is the vector of regression coefficients, X is a matrix of linguistic representations and Y_i a vector of perceptual feature i for the set of concrete nouns.

We now investigate ways in which the (quasi-) perceptual representations acquired via these methods can be combined with linguistic representations.

4.3 Information Combination

Canonical Correlation Analysis Canonical correlation analysis (CCA) (Hardoon et al., 2004) is an established statistical method for exploring relationships between two sets of random variables. The method determines a linear transformation of the space spanned by each of the sets of variables, such that the correlations between the sets of transformed variables is maximized.

Silberer and Lapata (2012) apply CCA in the present context of information fusion, with one set of random variables corresponding to perceptual features and another corresponding to linguistic features. Applied in this way, CCA provides a mechanism for reducing the dimensionality of the linguistic and perceptual representations such that the important interactions between them are preserved.[8] The transformed linguistic and perceptual vectors are then concatenated. We follow Silberer and Lapata by applying a kernalized variant of CCA.[9]

[7] Because the POS-tag and grammatical features are different for nouns and for verbs, we exclude them from our linguistic representations when implementing ridge regression.

[8] Dimensionality reduction is desirable in the present context because of the sparsity of our perceptual representations.

[9] The KernelCCA package in Python: http://pythonhosted.org/apgl/KernelCCA.html

Weighted Gram Matrix Combination The method we propose as an alternative means of fusing linguistic and extra-linguistic information is *weighted gram matrix combination*, which derives from an information combination technique applied to verb clustering by Reichart and Korhonen (2013). For a set of concepts $C = \{c_1, \ldots, c_n\}$ with representations $\{r_1, \ldots, r_n\}$, the method involves creating an $n \times n$ *weighted gram matrix* L in which

$$L_{ij} = S(r_i, r_j) \cdot \phi(r_i) \cdot \phi(r_j).$$

Here, S is again a similarity function (we use cosine similarity), and $\phi(r)$ is the *quality score* of r.

The quality scoring function ϕ can be any mapping $\mathbb{R}^n \rightarrow \mathbb{R}$ that reflects the importance of a concept relative to other concepts in C. In the present context, we follow Reichart and Korhonen (2013) in defining a quality score ϕ as the average cosine similarity of a concept with all other concepts in C

$$\phi(r_j) = \frac{1}{n} \sum_{i=1}^{n} S(r_i, r_j).$$

For $c_j \in C$, the matrix L then encodes a scalar projection of r_j onto the other members $r_{i \leq n}$, weighted by their quality. Each word representation in the set is thus mapped into a new space of dimension n determined by the concepts in C.

Converting concept representations to weighted gram matrix form has several advantages in the present context. First, both when evaluating and applying semantic representations, we generally require models to determine relations between concepts relative to others. We might, for instance, require close associates of a given word, a selection of potential synonyms, or the two most similar search queries in a given set. This relative nature of semantics is reflected by projecting representations into a space defined by the set of concepts themselves, rather than low-level features. It is also captured by the quality weighting, which lends primacy to concept dimensions that are central to the space.

Second, mapping representations of different dimension into vector spaces of equal dimension results in dense representations of equal dimension for each modality. This naturally lends equal weighting or status to each modality and resolves any issues

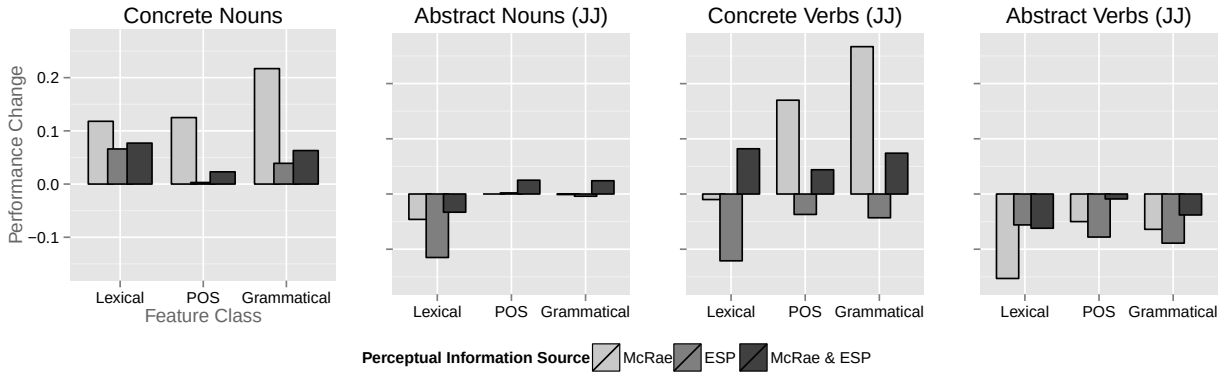

Figure 2: Additive change in Spearman's ρ when representations acquired from particular classes of linguistic features are combined with (actual or inferred) perceptual representations. Perceptual representations are derived from either the McRae Dataset, the ESP-Game Dataset or both (concatenated). For concepts other than concrete nouns, perceptual information is propagated using the Johns and Jones (JJ) method, and combined with simple concatenation.

of representations sparsity. In addition, the dimension equality in particular enables a wider range of mathematical operations for combining information sources. Here, we follow Reichart and Korhonen (2013) in taking the product of the linguistic and perceptual weighted gram matrices L and P, producing a new matrix containing fused representations for each concept

$$M = LPPL.$$

By taking the composite product $LPPL$ rather than LP or PL, M is symmetric and no ad hoc status is conferred to one modality over the other.

4.4 Results

The experiments in this section were designed to address the three questions specified in Section 1: (1) Which information sources are important for acquiring word concepts of different types? (2) Can perceptual information be propagated from concrete to abstract concepts? (3) What is the best way to combine the information from the different sources?

Question (1) To build on insights from Section 3, we first examined how perceptual input interacts with the three classes of linguistic features defined there. Figure 2 shows the additive difference in correlation between (i) models in which perceptual and particular linguistic features are concatenated and (ii) models based on just the linguistic features.

For concrete nouns and concrete verbs, (actual or inferred) perceptual information was beneficial in almost all cases. The largest improvement for both concept types was over grammatical features, achieved by including only the McRae data. This signals from this perceptual input and the grammatical features clearly reflect complementary aspects of the meaning of these concepts. We hypothesize that grammatical features (and POS features, which also perform strongly in this combination) confer information to concrete representations about the function and mutual interaction of concepts (the most 'relational' aspects of their meaning (Gentner, 1978)) which complements the more intrinsic properties conferred by perceptual features.

For abstract concepts, it is perhaps unsurprising that the overall contribution of perceptual information was smaller. Indeed, combining linguistic and perceptual information actually harmed performance on abstract verbs in all cases. For these concepts, the inferred perceptual features seem to obscure or contradict some of the information conveyed in the linguistic representations.

While the McRae data was clearly the most valuable source of perceptual input for concrete nouns and concrete verbs, for abstract nouns the combination of ESP-Game and McRae data was most informative. Both inspection of the data and cognitive theories (Rosch et al., 1976) suggest that entities identified in scenes, as in the ESP-Game dataset, generally correspond to a particular (basic) level of

Model	All Nouns	Conc. Nouns†	Abs. Nouns	All Verbs	Conc. Verbs	Abs. Verbs
Linguistic	0.175 — 0.335	0.169 — 0.317	0.233 — **0.344**	0.148 — 0.178	0.204 — 0.191	0.094 — **0.330**
(JJ)+Concat	0.116 — 0.375	0.258 — 0.442	0.190 — 0.267	0.129 — 0.162	**0.301** — 0.062	0.019 — 0.280
(JJ)+CCA	0.082 — 0.021	0.001 — 0.067	0.085 — -0.018	0.027— 0.213	0.079 — 0.276	**0.095** — 0.200
(JJ)+WGM	0.098 — 0.213	**0.397** — **0.523**	0.238 — 0.329	0.059 — 0.169	0.253 — 0.064	-0.080 — 0.254
RR+Concat	**0.232** — **0.432**		**0.248** — 0.343	0.013 — **0.212**	0.046 — **0.484**	0.023 — 0.133
RR+CCA	0.033 — -0.045		0.044 — -0.023	0.001 — -0.006	0.018 — 0.344	0.018 — 0.085
RR+WGM	0.094 — 0.069		0.232 — 0.327	**0.159** — 0.131	0.244 — 0.194	0.075 — 0.283
LR+	0.216 — 0.402		0.216 — 0.282	0.004 — 0.051	-0.051 — 0.139	-0.008 — 0.197

Table 4: Performance of different methods of information propagation (JJ = Johns and Jones, RR = ridge regression, LR = linear regression) and combination (Concat = concatenation, CCA = canonical correlation analysis, WGM = weighted gram matrix multiplication) across evaluation sets. Values are Spearman's ρ correlation with USF scores (left hand side of columns) and WordNet path similarity (right hand side). For the LR baseline we only report the highest score across the three combination types. †No propagation takes place for concrete nouns; this column reflects the performance of combination methods only.

the conceptual hierarchy. The ESP-Game data reflects relations between these basic-level concepts in the world, whereas the McRae data typically describes their (intrinsic) properties. Together, these sources seem to combine information on the properties of, and relations between, concepts in a way that particularly facilitates the learning of abstract nouns.

Question (2) The performance of different methods of information propagation and combination is presented in Table 4. The underlying linguistic representations in this case contained all three distributional feature classes. For more robust conclusions, in addition to the USF gold-standard we also measured the correlation between model output and the WordNet *path similarity* of words in our evaluation pairs. The path similarity between words w_1 and w_2 is the shortest distance between synsets of w_1 and w_2 in the WordNet taxonomy (Fellbaum, 1999), which correlates significantly with human judgements of concept similarity (Pedersen et al., 2004).[10]

The correlations with the USF data (left hand column, Table 4) of our linguistic-only models ($\rho = 0.094 - 0.233$) and best performing multi-modal models (on both concrete nouns, $\rho = 0.397$, and more abstract concepts, $\rho = 0.095 - 0.301$) were higher than the best comparable models described elsewhere (Feng and Lapata, 2010; Silberer and Lapata, 2012; Silberer et al., 2013).[11] This confirms both that the underlying linguistic space is of high quality and that the ESP and McRae perceptual input is similarly or more informative than the input applied in previous work.

Consistent with previous studies, adding perceptual input improved the quality of concrete noun representations as measured against both USF and path similarity gold-standards. Further, effective information propagation was indeed possible for both abstract nouns (USF evaluation) and concrete verbs (both evaluations). Interestingly, however, this was not the case for abstract verbs, for which no mix of propagation and combination methods produced an improvement on the linguistic-only model on either evaluation set. Indeed, as shown in Figure 2, no type of perceptual input generated an improvement in abstract verb representations, regardless of the underlying class of linguistic features.

This result underlines the link between concreteness, cognition and perception proposed in the psychological literature. More practically, it shows that concreteness can determine if propagation of perceptual input will be effective and, if so, the potential degree of improvement over text-only models.

Turning to means of propagation, both the Johns and Jones method and ridge regression outperformed the linear regression baseline on the majority of concept types in our evaluation. Across the five sets and ten evaluations on which propagation

[10]Other widely-used evaluation gold-standards, such as *WordSim 353* and the *MEN* dataset, do not contain a sufficient number of abstract concepts for the current purpose.

[11]Feng and Lapata (2010) report $\rho = .08$ for language-only and .12 for multi-modal models evaluated on USF over concrete and abstract concepts. Silberer and Lapata (2012) report $\rho = .14$ (language-only) and .35 (multi-modal) over concrete nouns.

takes place (*All Nouns, Abstract Nouns, All Verbs, Abstract Verbs* and *Concrete Verbs*), ridge regression performed more robustly, achieving the best performance on six evaluation sets compared to two for the Johns and Jones method.[12]

Question (3) Weighted gram matrix multiplication ($\rho = 0.397$ on USF and $\rho = 0.523$ on path similarity) outperformed both simple vector concatenation ($\rho = 0.258$ and $\rho = 0.442$) and CCA ($\rho = 0.001$ and $\rho = 0.067$) on concrete nouns. In the case of both abstract nouns and concrete verbs, however, the most effective means of combining quasi-perceptual information with linguistic representations was concatenation (abstract nouns, $\rho = 0.248$ and $\rho = 0.343$, concrete verbs, $\rho = 0.301$ and $\rho = 0.484$). One evident drawback of multiplicative methods such as weighted gram matrix combination is the greater inter-dependence of the information sources; a weak signal from one modality can undermine the contribution of the other modality. We hypothesize that this underlines the comparatively poor performance of the method on verbs and abstract nouns, as the perceptual input for concrete nouns is clearly a richer information source than the propagated features of more abstract concepts.

5 Conclusion

Motivated by the inherent difference between abstract and concrete concepts and the observation that abstract words occur more frequently in language, in this paper we have addressed the question of whether multi-modal models can enhance semantic representations of both concept types.

In Section 3, we demonstrated that different information sources are important for acquiring concrete and abstract noun and verb concepts. Within the linguistic modality, while lexical features are informative for all concept types, syntactic features are only significantly informative for abstract concepts.

In contrast, in Section 4 we observed that perceptual input is a more valuable information source for concrete concepts than abstract concepts. Nevertheless, perceptual input can be effectively propagated from concrete nouns to enhance representations of both abstract nouns and concrete verbs. Indeed, conceptual concreteness appears to determine the degree to which perceptual input is beneficial, since representations of abstract verbs, the most abstract concepts in our experiments, were actually degraded by this additional information. One important contribution of this work is therefore an insight into when multi-modal models should or should not aim to combine and/or propagate perceptual input to ensure that optimal representations are learned. In this respect, our conclusions align with the findings of Kiela and Hill (2014), who take an explicitly visual approach to resolving the same question.

Various methods for propagating and combining perceptual information with linguistic input were presented. We proposed ridge regression for inferring perceptual representations for abstract concepts, which proved more robust than alternatives across the range of concept types. This approach is particularly simple to implement, since it is based on an established statistical prodedure. In addition, we introduced weighted gram matrix combination for combining representations from distinct modalities of differing sparsity and dimension. This method produces the highest quality composite representations for concrete nouns, where both modalities represent high quality information sources.

Overall, our results demonstrate that the potential practical benefits of multi-modal models extend beyond concrete domains into a significant proportion of the lexical concepts found in language. In future work we aim to extend our experiments to concept types such as adjectives and adverbs, and to develop models that further improve the propagation and combination of extra-linguistic input.

Moreover, while we cannot draw definitive conclusions about human language processing, the effectiveness of the methods presented in this paper offer tentative support for the idea that even abstract concepts are *grounded* in the perceptual system (Barsalou et al., 2003). As such, it may be that, even in the more abstract cases of human communication, we find ways to see what people mean precisely by finding ways to see what they mean.

Acknowledgements

We thank The Royal Society and St John's College for their support.

[12] For these comparisons, the optimal combination method is selected in each case.

References

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

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.

Jeffrey R Binder, Chris F Westbury, Kristen A McKiernan, Edward T Possing, and David A Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17(6):905–917.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

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*, pages 248–255. IEEE.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906. Association for Computational Linguistics.

Christiane Fellbaum. 1999. *WordNet*. Wiley Online Library.

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. Association for Computational Linguistics.

Vittorio Gallese and George Lakoff. 2005. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3-4):455–479.

Dedre Gentner and Arthur B Markman. 1997. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Dedre Gentner. 1978. On relational meaning: The acquisition of verb meaning. *Child development*, pages 988–998.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics*, pages 241–247. Association for Computational Linguistics.

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

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Zellig Harris. 1954. Distributional structure. *Word*, 10(23):146–162.

Felix Hill, Anna Korhonen, and Christian Bentz. 2013. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Sabine Schulte Im Walde. 2006. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194.

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

Colin Kelly, Barry Devereux, and Anna Korhonen. 2010. Acquiring human-like feature-based conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, pages 61–69. Association for Computational Linguistics.

Douwe Kiela and Felix Hill. 2014. Improving multimodal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL 2014, Baltimore*. Association for Computational Linguistics.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Thomas K Landauer and Susan 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.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Chee Wee Leong and Rada Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *IJCNLP*, pages 1403–1407.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

Arthur B Markman and Edward J Wisniewski. 1997. Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1).

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

Raymond H Myers. 1990. *Classical and modern regression with applications*, volume 2. Duxbury Press Belmont, CA.

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.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 45(3):255.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics.

Roi Reichart and Anna Korhonen. 2013. Improved lexical acquisition through dpp-based verb clustering. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA, October. Association for Computational Linguistics.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

Gideon Rosen. 2001. Nominalism, naturalism, epistemic relativism. *Noûs*, 35(s15):69–91.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stockholm.

Paula J Schwanenflugel and Edward J Shoben. 1983. Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1):82.

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. Association for Computational Linguistics.

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*.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Tim Van de Cruys, Laura Rimell, Thierry Poibeau, Anna Korhonen, et al. 2012. Multiway tensor factorization for unsupervised lexical acquisition. *COLING 2012: Technical Papers*, pages 2703–2720.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 319–326. ACM.