Literal and Metaphorical Senses in Compositional Distributional Semantic Models

E. Darío Gutiérrez[1] **Ekaterina Shutova**[2] **Tyler Marghetis**[3] **Benjamin K. Bergen**[4]

[1] University of California San Diego
[2] University of Cambridge
[3] Indiana University Bloomington

edg@icsi.berkeley.edu tmarghet@cogsci.ucsd.edu
es407@cam.ac.uk bkbergen@ucsd.edu

Abstract

Metaphorical expressions are pervasive in natural language and pose a substantial challenge for computational semantics. The inherent compositionality of metaphor makes it an important test case for compositional distributional semantic models (CDSMs). This paper is the first to investigate whether metaphorical composition warrants a distinct treatment in the CDSM framework. We propose a method to learn metaphors as linear transformations in a vector space and find that, across a variety of semantic domains, explicitly modeling metaphor improves the resulting semantic representations. We then use these representations in a metaphor identification task, achieving a high performance of 0.82 in terms of F-score.

1 Introduction

An extensive body of behavioral and corpus-linguistic studies suggests that metaphors are pervasive in everyday language (Cameron, 2003; Steen et al., 2010) and play an important role in how humans define and understand the world. According to Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1981), individual metaphorical expressions, or *linguistic metaphors* (LMs), are instantiations of broader generalizations referred to as *conceptual metaphors* (CMs). For example, the phrases *half-baked idea*, *food for thought*, and *spoon-fed information* are LMs that instantiate the CM IDEAS ARE FOOD. These phrases reflect a mapping from the *source domain* of FOOD to the *target domain* of IDEAS (Lakoff, 1989). Two central claims of the CMT are that this mapping is systematic, in the sense that it consists of a fixed set of ontological correspondences, such as *thinking is preparing*, *communication is feeding*, *understanding is digestion*; and that this mapping can be productively extended to produce novel LMs that obey these correspondences.

Recent years have seen the rise of statistical techniques for metaphor detection. Several of these techniques leverage distributional statistics and vector-space models of meaning to classify utterances as literal or metaphorical (Utsumi, 2006; Shutova et al., 2010; Hovy et al., 2013; Tsvetkov et al., 2014). An important insight of these studies is that metaphorical meaning is not merely a property of individual words, but rather arises through cross-domain composition. The meaning of *sweet*, for instance, is not intrinsically metaphorical. Yet this word may exhibit a range of metaphorical meanings—e.g., *sweet dreams, sweet person, sweet victory*–that are created through the interplay of source and target domains. If metaphor is compositional, how do we represent it, and how can we use it in a compositional framework for meaning?

Compositional distributional semantic models (CDSMs) provide a compact model of compositionality that produces vector representations of phrases while avoiding the sparsity and storage issues associated with storing vectors for each phrase in a language explicitly. One of the most popular CDSM frameworks (Baroni and Zamparelli, 2010; Guevara, 2010; Coecke et al., 2010) represents nouns as vectors, adjectives as matrices that act on the noun vectors, and transitive verbs as third-order tensors that act on noun or noun phrase vectors. The meaning of a phrase is then derived by composing these lexical representations. The vast majority of such models build a single representation for all senses of a word, collapsing distinct senses together. One exception is the work of Kartsaklis and Sadrzadeh (2013a), who investigated homonymy, in which lexical items have identical form but unrelated meanings (e.g., *bank*). They found that deriving verb tensors from all instances of a homonymous form (as compared to training a separate tensor for each distinct sense) loses information and degrades the resultant phrase vector representations. To the best of our knowledge, there has not yet been a study of regular polysemy (i.e. metaphorical or metonymic sense distinctions) in the context of compositional distributional semantics. Yet, due to systematicity in metaphorical cross-domain mappings, there are likely to be systematic contextual sense distinctions that can be captured by a CDSM, improving the resulting semantic representations.

In this paper, we investigate whether metaphor, as a case of regular polysemy, warrants distinct treatment under a compositional distributional semantic framework. We propose a new approach to CDSMs, in which metaphorical meanings are distinct but structurally related to literal meanings. We then extend the generalizability of our approach by proposing a method to automatically learn metaphorical mappings as linear transformations in a CDSM. We focus on modeling adjective senses and evaluate our methods on a new data set of 8592 adjective-noun pairs annotated for metaphoricity, which we will make publicly available. Finally, we apply our models to classify unseen adjective-noun (AN) phrases as literal or metaphorical and obtain state-of-the-art performance in the metaphor identification task.

2 Background & Related Work

Metaphors as Morphisms. The idea of metaphor as a systematic mapping has been formalized in the framework of category theory (Goguen, 1999; Kuhn and Frank, 1991). In category theory, morphisms are transformations from one object to another that preserve some essential structure of the original object. Category theory provides a general formalism for analyzing relationships as morphisms in a wide range of systems (see Spivak (2014)). Category theory has been used to formalize the CM hypothesis with applications to user interfaces, poetry, and information visualization (Kuhn and Frank, 1991; Goguen and Harrell, 2010; Goguen and Harrell, 2005). Although these formal treatments of metaphors as morphisms are rigorous and well-formalized, they have been applied at a relatively limited scale. This is because this work does not suggest a straightforward and data-driven way to quantify semantic domains or morphisms, but rather focuses on the transformations and relations between semantic domains and morphisms, assuming some appropriate quantification has already been established. In contrast, our methods can learn representations of source-target domain mappings from corpus data, and so are inherently more scalable.

Compositional DSMs. Similar issues arose in modeling compositional semantics. Formal semantics has dealt with compositional meaning for decades, by using mathematical structures from abstract algebra, logic, and category theory (Montague, 1970; Partee, 1994; Lambek, 1999). However, formal semantics requires manual crafting of features. The central insight of CDSMs is to model the composition of words as algebraic operations on their vector representations, as provided by a conventional DSM (Mitchell and Lapata, 2008). Guevara (2010) and Baroni and Zamparelli (2010) were the first to treat adjectives and verbs differently from nouns. In their models, adjectives are represented by matrices that act on noun vectors. Adjective matrices can be learned using regression techniques. Other CDSMs have also been proposed and successfully applied to tasks such as sentiment analysis and paraphrase (Socher et al., 2011; Socher et al., 2012; Tsubaki et al., 2013; Turney, 2013).

Handling Polysemy in CDSMs. Several researchers argue that terms with ambiguous senses can be handled by DSMs without any recourse to additional disambiguation steps, as long as contextual information is available (Boleda et al., 2012; Erk and Padó, 2010; Pantel and Lin, 2002; Schütze, 1998; Tsubaki et al., 2013). Baroni et al. (2014) conjecture that CDSMs might largely avoid problems handling adjectives with multiple senses because the matrices for adjectives implicitly incorporate contextual information. However, they do draw a distinction between two ways in which the meaning of a term can vary. Continuous *polysemy*—the subtle and continuous variations in meaning resulting from the different contexts in which a word appears—is relatively tractable, in their opinion. This contrasts with discrete *homonymy*—the association of a single term with completely independent meanings (e.g., *light house* vs. *light work*). Baroni et al. concede that homonymy is more difficult to handle in

CDSMs. Unfortunately, they do not propose a definite way to determine whether any given variation in meaning is polysemy or homonymy, and offer no account of regular polysemy (i.e., metaphor and metonymy) or whether it would pose similar problems as homonymy for CDSMs.

To handle the problematic case of homonymy, Kartsaklis and Sadrzadeh (2013b) adapt a clustering technique to disambiguate the senses of verbs, and then train separate tensors for each sense, using the previously mentioned CDSM framework of Coecke et al. (2010). They found that prior disambiguation resulted in semantic similarity measures that correlated more closely with human judgments.

In principle, metaphor, as a type of regular polysemy, is different from the sort of semantic ambiguity described above. General ambiguity or vagueness in meaning (e.g. *bright light* vs *bright color*) is generally context-dependent in an unsystematic manner. In contrast, in regular polysemy meaning transfer happens in a systematic way (e.g. *bright light* vs. *bright idea*), which can be explicitly modeled within a CDSM. The above CDSMs provide no account of such systematic polysemy, which is the gap this paper aims to fill.

Computational Work on Metaphor. There is now an extensive literature on statistical approaches to metaphor detection. The investigated methods include clustering (Birke and Sarkar, 2006; Shutova et al., 2010; Li and Sporleder, 2010); topic modeling (Bethard et al., 2009; Li et al., 2010; Heintz et al., 2013); topical structure and imageability analysis (Strzalkowski et al., 2013); semantic similarity graphs (Sporleder and Li, 2009), and feature-based classifiers (Gedigian et al., 2006; Li and Sporleder, 2009; Turney et al., 2011; Dunn, 2013a; Dunn, 2013b; Hovy et al., 2013; Mohler et al., 2013; Neuman et al., 2013; Tsvetkov et al., 2013; Tsvetkov et al., 2014). We refer readers to the survey by Shutova (2015) for a more thorough review.

Most relevant to the present work are approaches that attempt to identify whether adjective-noun phrases are metaphorical or literal. Krishnakumaran and Zhu (2007) use AN co-occurrence counts and WordNet hyponym/hypernym relations for this task. If the noun and its hyponyms/hypernyms do not occur frequently with the given adjective, then the AN phrase is labeled as metaphorical. Krishnakumaran and Zhu's system achieves a precision of 0.67. Turney et al. (2011) classify verb and adjective phrases based on their level of concreteness or abstractness in relation to the noun they appear with. They learn concreteness rankings for words automatically (starting from a set of examples) and then search for expressions where a concrete adjective or verb is used with an abstract noun (e.g., *dark humor* is tagged as a metaphor; *dark hair* is not). They measure performance on a set of 100 phrases involving one of five adjectives, attaining an average accuracy of 0.79. Tsvetkov et al. (2014) train a random-forest classifier using several features, including abstractness and imageability rankings, WordNet supersenses, and DSM vectors. They report an accuracy of 0.81 on the Turney et al. (2011) AN phrase set. They also introduce a new set of 200 AN phrases, on which they measure an F-score of 0.85.

3 Experimental Data

Corpus. We trained our DSMs from a corpus of 4.58 billion tokens. Our corpus construction procedure is modeled on that of Baroni and Zamparelli (2010). The corpus consisted of a 2011 dump of English Wikipedia, the UKWaC (Baroni et al., 2009), the BNC (BNC Consortium, 2007), and the English Gigaword corpus (Graff et al., 2003). The corpus was tokenized, lemmatized, and POS-tagged using the NLTK toolkit (Bird and Loper, 2004) for Python.

Metaphor Annotations. We created an annotated dataset of 8592 AN phrases (3991 literal, 4601 metaphorical). Our choice of adjectives was inspired by the test set of Tsvetkov et al. (2014), though our annotated dataset is considerably larger. We focused on 23 adjectives that can have both metaphorical and literal senses, and which function as source-domain words in relatively productive CMs: TEMPERATURE (*cold*, *heated*, *icy*, *warm*), LIGHT (*bright*, *brilliant*, *dim*), TEXTURE (*rough*, *smooth*, *soft*); SUBSTANCE (*dense*, *heavy*, *solid*), CLARITY (*clean*, *clear*, *murky*), TASTE (*bitter*, *sour*, *sweet*), STRENGTH (*strong*, *weak*), and DEPTH (*deep*, *shallow*). We extracted all AN phrases involving these adjectives that occur in our corpus at least 10 times. We filtered out all phrases that require wider context to establish their meaning or metaphoricity—e.g., *bright side*, *weak point*.

The remaining phrases were annotated using a procedure based on Shutova et al. (2010). Annotators were encouraged to rely on their own intuition of metaphor, but were provided with the following guidance:

- For each phrase, establish the meaning of the adjective in the context of the phrase.
- Try to imagine a more basic meaning of this adjective in other contexts. Basic meanings tend to be: more concrete; related to embodied actions/perceptions/sensations; more precise; historically older/more "original".
- If you can establish a basic meaning distinct from the meaning of the adjective in this context, it is likely to be used metaphorically.

If requested, a randomly sampled sentence from the corpus that contained the phrase in question was also provided. The annotation was performed by one of the authors. The author's annotations were compared against those of a university graduate native English-speaking volunteer who was not involved in the research, on a sample of 500 phrases. Interannotator reliability (Cohen, 1960; Fleiss et al., 1969) was $\kappa = 0.80$ ($SE = .02$). Our annotated data set is publicly available at `http://bit.ly/1TQ5czN`

4 Representing Metaphorical Senses in a Compositional DSM

In this section we test whether separate treatment of literal and metaphorical senses is justified in a CDSM framework. In that case, training adjective matrix representations on literal and metaphorical subsets separately may result in systematically improved phrase vector representations, despite each matrix making use of fewer training examples.

4.1 Method

Our goal is to learn accurate vector representations for unseen adjective-noun (AN) phrases, where adjectives can take on metaphorical or literal senses. Our models build off the CDSM framework of Baroni and Zamparelli (2010), as extended by Li et al. (2014). Each adjective a is treated as a linear map from nouns to AN phrases:

$$\mathbf{p} = \mathbf{A}_a \mathbf{n},$$

where $\mathbf{p}$ is a vector for the phrase, $\mathbf{n}$ is a vector for the noun, and $\mathbf{A}_a$ is a matrix for the adjective.

Contextual Variation Model. The traditional representations do not account for the differences in meaning of an adjective in literal vs metaphorical phrases. Their assumption is that the contextual variations in meaning that are encoded by literal and metaphorical senses may be subtle enough that they can be handled by a single catch-all matrix per adjective, $\mathbf{A}_{\text{BOTH}(a)}$. In this model, every phrase i can be represented by

$$\mathbf{p}_i = \mathbf{A}_{\text{BOTH}(a)} \mathbf{n}_i \tag{1}$$

regardless of whether a is used metaphorically or literally in i. This model has the advantage of simplicity and requires no information about whether an adjective is being used literally or metaphorically. In fact, to our knowledge, all previous literature has handled metaphor in this way.

Discrete Polysemy Model Alternatively, the metaphorical and literal senses of an adjective may be distinct enough that averaging the two senses together in a single adjective matrix produces representations that are not well-suited for either metaphorical or literal phrases. Thus, the literal-metaphorical distinction could be problematic for CDSMs in the way that Baroni et al. (2014) suggested that homonyms are. Just as Kartsaklis and Sadrzadeh (2013a) solve this problem by representing each sense of a homonym by a different adjective matrix, we represent literal and metaphorical senses by different adjective matrices. Each literal phrase i is represented by

$$\mathbf{p}_i = \mathbf{A}_{\text{LIT}(a)} \mathbf{n}_i, \tag{2}$$

where $\mathbf{A}_{\text{LIT}(a)}$ is the literal matrix for adjective a. Likewise, a metaphorical phrase is represented by

$$\mathbf{p}_i = \mathbf{A}_{\text{MET}(a)} \mathbf{n}_i, \tag{3}$$

where $\mathbf{A}_{\text{MET}(a)}$ is the metaphorical matrix for a.

Learning. Given a data set of noun and phrase vectors $\mathcal{D}(a) = \{(\mathbf{n}_i, \mathbf{p}_i)\}_{i=1}^N$ for AN phrases involving adjective a extracted using a conventional DSM, our goal is to learn $\mathbf{A}_{\mathcal{D}(a)}$. This can be treated as an optimization problem, of learning an estimate $\hat{\mathbf{A}}_{\mathcal{D}(a)}$ that minimizes a specified loss function. In the case of the squared error loss, $L(\mathbf{A}_{\mathcal{D}(a)}) = \sum_{i \in \mathcal{D}(a)} \|\mathbf{p}_i - \mathbf{A}_{\mathcal{D}(a)} \mathbf{n}_i\|_2^2$, the optimal solution can be found precisely using ordinary least-squares regression. However, this may result in overfitting because of the large number of parameters relative to the number of samples (i.e., phrases). Regularization parameters $\lambda = (\lambda_1, \lambda_2)$ can be introduced to keep $\hat{\mathbf{A}}_{\mathcal{D}(a)}$ small:

$$\sum_{i \in \mathcal{D}(a)} \|\mathbf{p}_i - \hat{\mathbf{A}}_{\mathcal{D}(a)} \mathbf{n}_i\|_2^2 + R(\lambda; \hat{\mathbf{A}}_{\mathcal{D}(a)}),$$

where $R(\lambda; \hat{\mathbf{A}}_{\mathcal{D}}) = \lambda_1 \|\hat{\mathbf{A}}_{\mathcal{D}}\|_1 + \lambda_2 \|\hat{\mathbf{A}}_{\mathcal{D}}\|_2$. This approach, known as elastic-net regression (Zou and Hastie, 2005), produces better adjective matrices than unregularized regression (Li et al., 2014). Note that the same procedure can be used to learn the adjective representations in both the Contextual Variation model and the Discrete Polysemy model by varying what phrases are included in the training set $\mathcal{D}(a)$. In the Contextual Variation model $\mathcal{D}(a)$ includes both metaphorical and literal phrases, while in the Discrete Polysemy model it includes only metaphorical phrases when learning $\hat{\mathbf{A}}_{\text{MET}(a)}$ and testing on metaphorical phrases (and only literal phrases when learning $\hat{\mathbf{A}}_{\text{LIT}(a)}$ and testing on literal phrases).

4.2 Experimental Setup

Extracting Noun & Phrase Vectors. Our approach for constructing term vector representations is similar to that of Dinu et al. (2013). We first selected the 10K most frequent nouns, adjectives, and verbs to serve as context terms. We then constructed a co-occurrence matrix that recorded term-context co-occurrence within a symmetric 5-word context window of the 50K most frequent POS-tagged terms in the corpus. We then used these co-occurrences to compute the positive pointwise mutual information (PPMI) between every pair of terms, and collected these into a term-term matrix. Next, we reduced the dimensionality of this matrix to 100 dimensions using singular-value decomposition. Additionally, we computed "ground truth" distributional vectors for all the annotated AN phrases in our data set by treating the phrases as single terms and computing their PPMI with the 50K single-word terms, and then projecting them onto the same 100-dimensional basis.

Training Adjective Matrices. For each adjective a that we are testing, we split the phrases involving that adjective into two subsets, the literal (LIT) subset and the metaphorical (MET) subset. We then split the subsets into 10 folds, so that we do not train and test any matrices on the same phrases. For each fold k, we train three adjective matrices: $\hat{\mathbf{A}}_{\text{MET}(a)}$ using all phrases from the MET set not in fold k; $\hat{\mathbf{A}}_{\text{LIT}(a)}$ using all phrases from the LIT set not in fold k; and $\hat{\mathbf{A}}_{\text{BOTH}(a)}$ using all the phrases from either subset not in fold k. Within each fold, we use nested cross-validation as outlined in Li et al. (2014) to determine the regularization parameters for each regression problem.

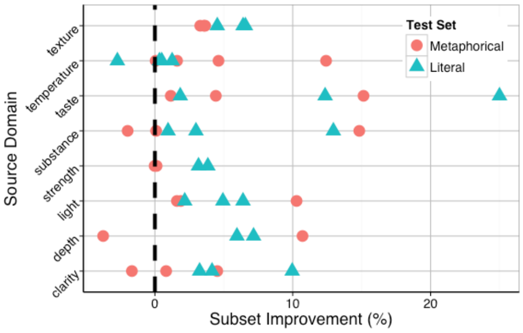

Figure 1: Reduction in error from training on targeted subset (MET/LIT) rather than on all phrases.

4.3 Evaluating Vector Representations

Evaluation. Our goal is to produce a vector prediction of each phrase that will be close to its ground truth distributional vector. Phrase vectors directly extracted from the corpus by treating the phrase as a single term are the gold standard for predicting human judgment and producing paraphrases (Dinu et al., 2013), so we use these as our ground truth. The quality of the vector prediction for phrase i is measured using the cosine distance between the phrase's ground truth vector $\mathbf{p}_i$ and the vector prediction $\hat{\mathbf{p}}_i$:

$$err(\hat{\mathbf{p}}_i) = 1 - \cos(\hat{\mathbf{p}}_i, \mathbf{p}_i).$$

We then analyze the benefit of training on a reduced subset by calculating a "subset improvement" (SI) score for the MET and LIT subsets of each adjective a. We define the SI for each subset $\mathcal{D}(a) \in \{\text{LIT}(a), \text{MET}(a)\}$ as:

$$SI(\mathcal{D}(a)) = 1 - \frac{\sum_{i \in \mathcal{D}(a)} err(\hat{\mathbf{A}}_{\mathcal{D}(a)} \mathbf{n}_i)}{\sum_{i \in \mathcal{D}(a)} err(\hat{\mathbf{A}}_{\text{BOTH}(a)} \mathbf{n}_i)}$$

Positive values of SI thus indicate improved performance when trained on a reduced subset compared to the full set of phrases. For example $SI_{\text{LIT}(a)} = 5\%$ tells us that predicting the phrase vectors for LIT phrases of adjective a using the LIT matrix resulted in a 5% reduction in mean cosine error compared to predicting the phrase vectors using the BOTH matrix.

Results. The results are summarized in Fig. 1. Each point indicates the SI for a single adjective and for a single subset. Adjectives are grouped by source domain along the y-axis. Overall, almost every item shows a subset improvement; and, for every source domain, the majority of adjectives show a subset improvement.

We analyzed per-adjective SI by fitting a linear mixed-effects model, with a fixed intercept, a fixed effect of test subset (MET vs. LIT), a random effect of source domain, and the maximal converging random effects structure (uncorrelated random intercepts and slopes) (Barr et al., 2013). Training on a targeted subset improved performance by $4.4\% \pm 0.009(SE)$ ($p = .002$). There was no evidence that this differed by test subset (i.e., metaphorical vs. literal senses, $p = .35$). The positive SI from training on a targeted subset suggests that metaphorical and literal uses of the same adjective are semantically distinct.

4.4 Metaphor Classification

Method. The results of the previous section suggest a straightforward classification rule: classify unseen phrase i involving adjective a as metaphorical if $\cos(\mathbf{p}_i, \hat{\mathbf{A}}_{\text{MET}(a)}\mathbf{n}_i) < \cos(\hat{\mathbf{A}}_{\text{LIT}(a)}\mathbf{n}_i)$. Otherwise, we classify it as literal.

Evaluation. We test this method on our data set of 8593 annotated AN phrases using 10-fold cross validation. It is possible that our method's classification performance is not due to the compositional aspect of the model, but rather to some semantic coherence property among the nouns in the AN phrases that we are testing. To control for this possibility, we compare the performance of our method against four baselines. The first baseline, NOUN-NN, measures the cosine distance between the vector for the noun of the AN phrase being tested and the noun vectors of the nouns participating in an AN phrase in the training folds. The test phrase is then assigned the label of the AN phrase whose noun vector is nearest. PHRASE-NN proceeds similarly, but using the ground-truth phrase vectors for the test phrase and the training phrases. The test phrase is then assigned the label of the AN phrase whose vector is nearest. The baseline NOUN-CENT first computes the centroid of the noun vectors of the training phrases that are literal, and the centroid of the noun vectors of the training phrases that are metaphorical. It then assigns the test phrase the label of the centroid whose cosine distance from the test phrase's noun vector is smallest. PHRASE-CENT, proceeds similarly, but using phrase vectors. We measure performance against the manual annotations.

Results. Our classification method achieved a held-out F-score of 0.817, recall of 0.793, precision of 0.842, and accuracy of 0.809. These results were superior to those of the baselines (Table 1). These results are competitive with the state of the art and demonstrate the importance of compositionality in metaphor identification.

Method	F-score	Precision	Recall	Accuracy
MET-LIT	0.817	0.842	0.793	0.809
NOUN-NN	0.709	0.748	0.675	0.703
PHRASE-NN	0.590	0.640	0.547	0.592
NOUN-CENT	0.717	0.741	0.695	0.706
PHRASE-CENT	0.629	0.574	0.695	0.559

Table 1: Performance of the method of §4.4 (MET-LIT) against various baselines.

5 Metaphors as Linear Transformations

One of the principal claims of the CM hypothesis is that CMs are productive: A CM (i.e., mapping) can generate endless new LMs (i.e., linguistic expressions). Cases where the LMs involve an adjective that has already been used metaphorically and for which we have annotated metaphorical and literal examples can be handled by the methods of §4, but when the novel LM involves an adjective that has only been observed in literal usage, we need a more elaborate model. According to the CM hypothesis, an adjective's metaphorical meaning is a result of the action of a source-to-target CM mapping on the adjective's literal sense. If so, then given an appropriate representation of this mapping it should be possible to infer the metaphorical sense of an adjective without ever seeing metaphorical exemplars—that is, using only the adjective's literal sense. Our next experiments seek to determine whether it is possible to represent and learn CM mappings as linear maps in distributional vector space.

5.1 Model

We model each CM mapping $\mathcal{M}$ from source to target domain as a linear transformation $\mathbf{C}_{\mathcal{M}}$:

$$\mathbf{A}_{\text{MET}(a)}\mathbf{n}_i \approx \mathbf{C}_{\mathcal{M}}\mathbf{A}_{\text{LIT}(a)}\mathbf{n}_i \qquad (4)$$

We can apply a two-step regression to learn $\mathbf{C}_{\mathcal{M}}$. First we apply elastic-net regression to learn the literal adjective matrix $\hat{\mathbf{A}}_{\text{LIT}(a)}$ as in §4.2. Then we can substitute this estimate into Eq. (4), and apply elastic-net regression to learn the $\hat{\mathbf{C}}_{\mathcal{M}}$ that minimizes the regularized squared error loss:

$$\sum_{a\in\mathcal{M}}\sum_{i\in\mathcal{D}(a)} \|\mathbf{p}_i - \hat{\mathbf{C}}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i\|_2^2 + R(\lambda; \hat{\mathbf{C}}_{\mathcal{M}}).$$

To learn $C_{\mathcal{M}}$ in this regression problem, we can pool together and train on phrases from many different adjectives that participate in $\mathcal{M}$.

5.2 Experimental Setup

We used a cross-validation scheme where we treated each adjective in a source domain as a fold in training the domain's metaphor transformation matrix. The nested cross-validation procedure we use to set regularization parameters λ and evaluate performance requires at least 3 adjectives in a source domain, so we evaluate on the 6 source domain classes containing at least 3 adjectives. The total number of phrases for these 19 adjectives is 6987 (3659 metaphorical, 3328 literal).

5.3 Evaluating Vector Representations

Evaluation. We wish to test whether CM mappings learned from one set of adjectives are transferable to new adjectives for which metaphorical phrases are unseen. As in §4, models were evaluated using cosine error compared to the ground truth phrase vector representation. Since our goal is to improve the vector representation of metaphorical phrases given no metaphorical annotations, we measure performance on the MET phrase subset for each adjective. We compare the performance of the transformed LIT matrix $\mathbf{C}_{\mathcal{M}}\mathbf{A}_{\text{LIT}(a)}$ against the performance of the original LIT matrix $\mathbf{A}_{\text{LIT}(a)}$ by defining the metaphor transformation improvement (MTI) as:

$$MTI(a) = 1 - \frac{\sum_{i \in \text{MET}} err(\mathbf{C}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a)})}{\sum_{i \in \text{MET}} err(\hat{\mathbf{A}}_{\text{LIT}(a)})}.$$

Results. Per-adjective MTI was analyzed with a linear mixed-effects model, with a fixed intercept, a random effect of source domain, and random intercepts. Transforming the LIT matrix using the CM mapping matrix improved performance by $11.5\% \pm 0.023(SE)$ ($p < .001$). On average, performance improved for 18 of 19 adjectives and for every source domain ($p = .03$, binomial test; Fig. 2). Thus, mapping structure is indeed shared across adjectives participating in the same CM.

5.4 Metaphor Classification

Method. Once again our results suggest a procedure for metaphor classification. This procedure can classify phrases involving adjectives without seeing any metaphorical annotations. For any unseen phrase i involving an adjective a_i, we classify the phrase as metaphorical if $\cos(\mathbf{p}_i, \hat{\mathbf{C}}_{\mathcal{M}}\hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i) < \cos(\mathbf{p}_i, \hat{\mathbf{A}}_{\text{LIT}(a_i)}\mathbf{n}_i)$. Otherwise, we classify it as literal. We used the same procedure as in §4.2 to learn $\hat{\mathbf{A}}_{\text{LIT}(a_i)}$.

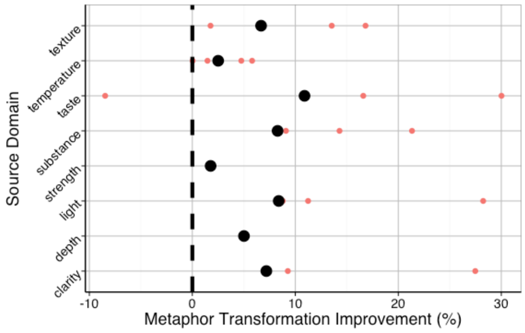

Figure 2: Reduction in error from transforming LIT matrix using metaphorical mapping. Mean change was positive for every domain (large black), and for all but one adjective (small red).

Method	F-score	Precision	Recall	Accuracy
TRANS-LIT	0.793	0.716	0.819	0.804
MET-LIT	0.838	0.856	0820	0.833
NOUN-NN	0.692	0.732	0.655	0.693
PHRASE-NN	0.575	0.625	0.532	0.587
NOUN-CENT	0.703	0.722	0.685	0.696
PHRASE-CENT	0.610	0.552	0.681	0.542

Table 2: Performance of method of §5.4 (TRANS-LIT) against method of §4.4 (MET-LIT) and various baselines.

Results. Our method achieved an F-score of 0.793 on the classification of phrases involving unseen adjectives. On this same set of phrases, the method of §4.4 achieved an F-score of 0.838. Once again, the performance of our method was superior to the performance of the baselines (Table 2; the MET-LIT figures in Table 2 differ slightly from those in Table 1 because only 19 of 23 adjectives are tested). For comparison, we also include the classification performance using the MET-LIT method of §4.4. While MET-LIT slightly outperforms TRANS-LIT, the latter has the benefit of not needing annotations for metaphorical phrases for the test adjective. Hence, our approach is generalizable to cases where such annotations are unavailable with only slight performance reduction.

6 Discussion

Overall, our results show that taking metaphor into account has the potential to improve CDSMs and expand their domain of applicability. The findings of §4 suggest that collapsing across metaphorical and literal uses may hurt accuracy of vector representations in CDSMs. While the method in §4 depends on explicit annotations of metaphorical and literal senses, the method in §5 provides a way to generalize these representations to adjectives for which metaphorical training data is unavailable, by showing that metaphorical mappings are transferable across adjectives from the same source domain. Note that an accurate matrix representation of the literal sense of each adjective is still required in the experimental setup of §5. This particular choice of setup allowed a proof of concept of the hypothesis that metaphors function as cross-domain transformations, but in principle it would be desirable to learn transformations from a general BOTH matrix representation for any adjective in a source domain to its MET matrix representation. This would enable improved vector representations of metaphorical AN phrases without annotation for unseen adjectives.

The success of our models on the metaphor classification tasks demonstrates that there is information about metaphoricity of a phrase inherent in the composition of the meanings of its components. Notably, our results show that this metaphorical compositionality can be captured from corpus-derived distributional statistics. We also noticed some trends at the level of individual phrases. In particular, classification performance and vector accuracy tended to be lower for metaphorical phrases whose nouns are distributionally similar to nouns that tend to participate in literal phrases (e.g., *reception* is similar to *foyer* and *refreshment* in our corpus; *warm reception* is metaphorical while *warm foyer* is literal). Another area where classification accuracy is low is in phrases with low corpus occurrence frequency. The ground truth vectors for these phrases exhibit high sample variance and sparsity. Many such phrases sound paradoxical (e.g., *bitter sweetness*).

Our results could also inform debates within cognitive science. First, cognitive scientists debate whether words that are used both literally and figuratively (e.g., *long road*, *long meeting*) are best understood as having a single, abstract meaning that varies with context or two distinct but related meanings. For instance, some argue that domains like space, time, and number operate over a shared, generalized magnitude system, yet others maintain that our mental representation of time and number is distinct from our mental representation of space, yet inherited metaphorically from it (Winter et al., 2015). Our results suggest that figurative and literal senses involve quite different patterns of use. This is statistical evidence that adjectives that are used metaphorically have distinct related senses, not a single abstract sense.

Second, the Conceptual Metaphor Theory account hypothesizes that LMs are an outgrowth of metaphorical thought, which is in turn an outgrowth of embodied experiences that conflate source and target domains—experience structures thought, and thought structures language (Lakoff, 1993). However, recent critics have argued for the opposite causal direction: Linguistic regularities may drive the mental mapping between source and target domains (Hutchinson and Louwerse, 2013; Casasanto, 2014; Hutchinson and Louwerse, 2014). Our results show that, at least for AN pairs, the semantic structure of a source domain and its mapping to a metaphorical target domain are available in the distributional statistics of language itself. There may be no need, therefore, to invoke embodied experience to explain the prevalence of metaphorical thought in adult language users. A lifetime of experience with literal and metaphorical language may suffice.

7 Conclusion

We have shown that modeling metaphor explicitly within a CDSM can improve the resulting vector representations. According to our results, the systematicity of metaphor can be exploited to learn linear transformations that represent the action of metaphorical mappings across many different adjectives in the same semantic domain. Our classification results suggest that the compositional distributional semantics of a phrase can inform classification of the phrase for metaphoricity.

Beyond improvements to the applications we presented, the principles underlying our methods also show potential for other tasks. For instance, the LIT and MET adjective matrices and the CM mapping matrix learned with our methods could be applied to improve automated paraphrasing of AN phrases. Our work is also directly extendable to other syntactic constructions. In the CDSM framework we apply, verbs would be represented as third-order tensors. Tractable and efficient methods for estimating these verb tensors are now available (Fried et al., 2015). It may also be possible to extend the coverage of our system by using automated word-sense disambiguation to bootstrap annotations and therefore construct LIT and MET matrices in a minimally supervised fashion (Kartsaklis et al., 2013b). Finally, it would be interesting to investigate modeling metaphorical mappings as nonlinear mappings within the deep learning framework.

Acknowledgments

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. Ekaterina Shutova's research is supported by the Leverhulme Trust Early Career Fellowship.

References

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.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Steven Bethard, Vicky Tzuyin Lai, and James H. Martin. 2009. Topic model analysis of metaphor frequency for psycholinguistic stimuli. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 9–16. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 1–4.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

BNC Consortium. 2007. British National Corpus, Version 3 BNC XML edition.

Gemma Boleda, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally. 2012. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1223–1233. Association for Computational Linguistics.

Lynne Cameron. 2003. *Metaphor in Educational Discourse*. A&C Black, London.

Daniel Casasanto. 2014. Development of metaphorical thinking: The role of language. In Mike Borkent, Barbara Dancygier, and Jennifer Hinnell, editors, *Language and the Creative Mind*, pages 3–18. CSLI Publications, Stanford.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. In *Linguistic Analysis (Lambek Festschrift)*, pages 345–384.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. educational and psychosocial measurement.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of the ACL 2013 Workshop on Continuous Vector Space Models and their Compositionality (CVSC 2013)*, pages 50–58, East Stroudsburg, Pennsylvania. ACL.

Jonathan Dunn. 2013a. Evaluating the premises and results of four metaphor identification systems. In *Computational Linguistics and Intelligent Text Processing*, pages 471–486. Springer.

Jonathan Dunn. 2013b. What metaphor identification systems can tell us about metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 92–97. Association for Computational Linguistics.

Joseph L. Fleiss, Jacob Cohen, and B.S. Everitt. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323.

Daniel Fried, Tamara Polajnar, and Stephen Clark. 2015. Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, Beijing.

Matt Gedigian, John Bryant, Srini Narayanan, and Branimir Ciric. 2006. Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, pages 41–48, New York. Association for Computational Linguistics.

Joseph A. Goguen and D. Fox Harrell. 2005. 7 information visualisation and semiotic morphisms. *Studies in Multidisciplinarity*, 2:83–97.

Joseph A. Goguen and D. Fox Harrell. 2010. Style: A computational and conceptual blending-based approach. In *The Structure of Style*, pages 291–316. Springer, New York.

Joseph Goguen. 1999. An introduction to algebraic semiotics, with application to user interface design. In *Computation for metaphors, analogy, and agents*, pages 242–291. Springer.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English Gigaword. *Linguistic Data Consortium, Philadelphia*.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Ilana Heintz, Ryan Gabbard, Mahesh Srinivasan, David Barner, Donald S Black, Marjorie Freedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphor with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

Sterling Hutchinson and Max Louwerse. 2013. Language statistics and individual differences in processing primary metaphors. *Cognitive Linguistics*, 24(4):667–687.

Sterling Hutchinson and Max M. Louwerse. 2014. Language statistics explain the spatial–numerical association of response codes. *Psychonomic Bulletin & Review*, 21(2):470–478.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, et al. 2013a. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1590–1601.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013b. Separating disambiguation from composition in distributional semantics. In *Proceedings of the 2013 Conference on Computational Natural Language Learning*, pages 114–123.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational approaches to Figurative Language*, pages 13–20. Association for Computational Linguistics.

Werner Kuhn and Andrew U Frank. 1991. A formalization of metaphors and image-schemas in user interfaces. In *Cognitive and linguistic aspects of geographic space*, pages 419–434. Springer.

George Lakoff and Mark Johnson. 1981. *Metaphors we live by*. University of Chicago Press, Chicago.

George Lakoff. 1989. Some empirical results about the nature of concepts. *Mind & Language*, 4(1-2):103–129.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, Cambridge.

Joachim Lambek. 1999. Type grammar revisited. In *Logical aspects of computational linguistics*, pages 1–27. Springer, Berlin.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 315–323. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2010. Using Gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300. Association for Computational Linguistics.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics.

Jiming Li, Marco Baroni, and Georgiana Dinu. 2014. Improving the lexical function composition model with pathwise optimized elastic-net regression. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 434–442.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL-08: HLT*, pages 236–244.

Michael Mohler, David Bracewell, David Hinote, and Marc Tomlinson. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35.

Richard Montague. 1970. English as a formal language. In B Visentini and et al, editors, *Linguaggi nella Società e nella Tecnica*. Edizioni di Comunitá, Milan.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PLoS ONE*, 8:e62343.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619. ACM.

Barbara H. Partee. 1994. Lexical semantics and compositionality. In Lila Gleitman and Mark Liberman, editors, *Invitation to Cognitive Science 2nd Edition, Part I: Language*. MIT Press, Cambridge, Mass., USA.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

Ekatrina Shutova. 2015. Design and evaluation of metaphor processing systems. volume Forthcoming.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

David I. Spivak. 2014. *Category Theory for the Sciences*. MIT Press, Cambridge, Mass., USA.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing, Amsterdam/Philadelphia.

Tomek Strzalkowski, George A. Broadwell, Sarah Taylor, Laurie Feldman, Boris Yamrom, Samira Shaikh, Ting Liu, Kit Cho, Umit Boz, Ignacio Cases, and Kyle Elliot. 2013. Robust extraction of metaphors from novel data. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 67–76, Atlanta, Georgia. Association for Computational Linguistics.

Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto. 2013. Modeling and learning semantic co-compositionality through prototype projections and neural networks. In *The 2013 Conference on Empirical Methods in Natural Language Processing*, pages 130–140.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.

Peter D. Turney. 2013. Distributional semantics beyond words: supervised learning of analogy and paraphrase. *Transactions of the Association for Computational Linguistics (TACL)*, 1:353–366.

Akira Utsumi. 2006. Computational exploration of metaphor comprehension processes. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society (CogSci2006)*, pages 2281–2286.

Bodo Winter, Tyler Marghetis, and Teenie Matlock. 2015. Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. *Cortex*, 64:209–224.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.