Modeling metaphor perception with distributional semantics vector space models

Kat R. Agres,[1] Stephen McGregor,[1] Karolina Rataj,[2,3]
Matthew Purver,[1] Geraint A. Wiggins[1]

[1] Queen Mary University of London, London, UK
{kathleen.agres,s.e.mcgregor,m.purver,geraint.wiggins}@qmul.ac.uk
[2] Department of Cognitive Psychology and Ergonomics,
University of Twente, Enschede, the Netherlands
[3] Faculty of English, Adam Mickiewicz University, Poznań, Poland
krataj@wa.amu.edu.pl

Abstract. In this paper, we present a novel application of a computational model of word meaning to capture human judgments of the linguistic properties of *metaphoricity*, *familiarity*, and *meaningfulness*. We present data gathered from human subjects regarding their ratings of these properties over a set of word pairs specifically designed to exhibit varying degrees of metaphoricity. We then investigate whether these properties can be measured in terms of geometric features of a model of distributional lexical semantics. We compare the performance of two models, our own Concept Discovery Model which dynamically constructs context-sensitive subspaces, and a state-of-the-art static distributional semantic model, and find that our dynamic model performs significantly better in its measurement of metaphoricity.

Keywords: metaphor, distributional semantics, vector space models, computational creativity

1 Introduction

In this study, we investigate whether computational models of lexical meaning might help explain human comprehension of metaphors. We examine several alternative models, based on language statistics in a large general collection of English, to see if they capture relations between words which correlate with human judgments of metaphoricity.

Psycholinguistic studies Research on metaphor in human participants has attempted to clarify the mechanisms underlying understanding metaphoric language. One of the earliest approaches, the standard pragmatic model, stipulates that the literal meaning of a metaphoric sentence needs to be rejected before the figurative meaning is generated [10]. Behavioral studies inspired by this model have shown that participants do not use more time to comprehend metaphoric than literal sentences, and that metaphoric meaning is generated in parallel

with the literal meaning of an utterance [8, 29]. These studies, however, did not make a distinction between conventional and novel metaphors. Later reports have shown that novel metaphors do require more processing time than literal sentences, while conventional metaphor and literal language comprehension take comparable time [2].

Recently, a number of electrophysiological (EEG) studies investigating metaphor comprehension have been reported in which the N400 component, a negative-going wave observed between 300 and 500ms after the presentation of the critical stimulus, has received considerable attention. Larger N400 amplitudes have been observed in the processing of metaphoric as compared to literal sentences, with no differences in component latency (i.e., the time window within which the effect is observed) or scalp distribution (i.e., the sites on the scalp over which the effect is present) (e.g., [2]). This increase in amplitude has been interpreted as reflecting more activity in memory needed to retrieve the semantic information necessary for comprehension of metaphoric as compared to literal sentences[16]. At the same time, comparable latency and scalp distribution of the component might be indicative of the involvement of similar mechanisms in literal and metaphoric language comprehension. Interestingly, differences have been found between conventional and novel metaphors, with the N400 amplitudes for conventional metaphors falling in-between those for novel metaphoric and literal utterances [2]. This graded effect has not been observed in reaction time studies, which demonstrates that ERP measures offer greater sensitivity to the time course of cognitive processes involved in metaphor comprehension. These findings raise important questions concerning the nature of the mechanisms involved in understanding metaphors.

One of the approaches that has attempted to elucidate metaphor comprehension mechanisms is the structure mapping model and its descendant, the career of metaphor model, which stipulate that the same mechanisms are involved in the comprehension of literal comparisons, similes, and metaphors [5, 30]. Within this view, comprehending metaphoric sentences, like *my mind is a warehouse*, requires a mapping that involves a symmetrical mechanism of alignment of relational commonalities in the source (*warehouse*) and target (*mind*), together with an asymmetrical mechanism of inference projection from the source to the target. Moreover, the career of metaphor model assumes that while novel metaphors are understood via comparison, categorization is involved in conventional metaphor understanding. These assumptions have received some support in ERP studies, which have also shown that a shared mapping process may be involved in categorization and comparison [9, 17]. Moreover, comparison seems to facilitate not only novel, but also conventional metaphor comprehension, although this facilitation is observed at later processing stages than in the case of novel metaphors.

Computational studies Computational approaches to metaphor have generally focused on a combination of pattern matching and hand coded information processing [25]. The KARMA model for metaphor understanding [7], for instance, attempts to encode environmentally grounded knowledge about action in the world into a framework of transferable domains. In a similar vein, ATT-Meta

[3] asks users to provide domain specific knowledge about entities and processes and then ports this knowledge between different contexts—this context sensitive aspect of the system in particular aligns with both the approach to theoretical work on metaphor and to the computational work presented here.

Moving towards data-driven approaches, a model for metaphor comprehension has been described that employs latent semantic analysis, a statistical technique for building spaces of word similarity, to the selection of the salient features trafficked between a metaphoric source and target [15]. This technique, along with a similar method involving the selection of transferred features based on proximity in a semantic space [27], bears comparison to our dynamically contextual method as described in Section 3.2: each of these models attempts to extract particular features of a semantic space in order to capture the semantic context of a metaphor. A more recent description of a Service-Oriented Architecture [28] discovers properties of source and target by matching patterns within large-scale web corpora and then looks for properties salient in the source which can be transferred to the target. Other contemporary approaches have tended towards the more overtly statistical, with for instance the application of linear algebraic operations to model the metaphoric composition of vector space type word representations [11].

The computational component of the work presented here broadly falls within the paradigm of the *distributional hypothesis*, which holds that "words that occur in similar contexts have similar meanings," [26, p. 148]. The general methodology of distributional semantics involves the traversal of large scale textual corpora in order to build spaces of word-vectors where the proximity of two vectors reflects the tendency for those two words to be observed co-occurring with similar terms [6]. Initial approaches to distributional semantics typically involved building up word representations based on straight-up co-occurrence counts [24], while more recent methodologies have often incorporated matrix factorisation techniques to derive dense matrices from co-occurrence statistics [13] or employed neural network architectures to derive *word embeddings* from observations of co-occurrences across iterative traversals of a corpus [4].

In the study presented here, we will in particular be comparing Word2Vec [21], a neural network driven model for generating word embeddings that has achieved state-of-the-art results on tests of word similarity and analogy completion, with our own Concept Discovery Model [19], which deploys a word-counting approach to distributional semantics to dynamically construct contextualised subspaces in which conceptual relationships play out as geometric relationships [20]. We will be examining the ways in which spaces generated by each model compare with human assessments of the degree of *metaphoricity*, *familiarity*, and *meaningfulness* in noun-verb word pairings. With this in mind, recent work using distributional models enriched with information from lexical and associative knowledge bases to build spaces of word-vectors constructed for detecting similarity or relatedness should be taken into consideration [14].

2 Modeling human metaphor judgements

Our objective in this study is to explore the ways in which geometric models of word meaning can capture the perception of metaphor, and in particular can measure the degree to which two-word phrases are perceived as being metaphorical. To do so, we compare words' relations in the geometric model with human judgments of metaphoricity, via a set of empirically-derived normative data. Note that while data were collected for three types of norming measures (metaphoricity, meaningfulness, and familiarity), the principal aim of the computational work is to model the perception of metaphoricity, that is, to discover meaningful subspaces that reflect the extent to which a two-term expression is perceived as being metaphorical.

2.1 Materials

The materials were collected for an ERP study, which investigated metaphor comprehension in bilinguals [12]. Verb-noun word dyads in Polish (native language) and English (second language) were used in the ERP experiment. In each case, the verb was considered the metaphoric source and the noun the target: so, for example, in the instance of the conventional metaphor "cut pollution", some salient property of the action CUTTING is being transferred to the entity POLLUTION.

Prior to the ERP experiment, five normative studies were carried out to ensure the word pairs fell within the following three categories: novel metaphors (e.g., *to harvest courage*), conventional metaphors (e.g., *to gather courage*), and literal expressions (e.g., *to experience courage*). Based on the results of the normative studies, the final set of 228 English verb-noun word dyads (76 in each category) was selected for the purpose of the current study. Out of the five normative studies, four will be reported here. The statistical analyses consisted of mixed-design analyses of variance (ANOVAs), with utterance type as a within-subject factor and survey block as a between-subject factor. No main effect of survey block was observed. Significance values for the pairwise comparisons were corrected for multiple comparisons using the Bonferroni correction. When Mauchlys tests showed that the assumption of sphericity was violated, the Greenhouse-Geisser correction was applied. In such cases, the original degrees of freedom are reported with the corrected p value. The demographic data for the participants of the four normative studies are presented in Table 1.

Table 1. Demographic characteristics of participants of the four normative studies, including the number of participants (number of female participants) and mean age.

Normative study type	**Number of participants(female)**	**Mean age**
Cloze probability	140 (65)	23
Meaningfulness ratings	133 (61)	22
Familiarity ratings	101 (55)	23
Metaphoricity ratings	102 (59)	22

Cloze probability Because reduced N400 amplitudes have been observed in relation to expected as compared to unexpected words, a cloze probability test was performed prior to the ERP study to ensure the second word in a given word dyad was not highly anticipated by the participants of the ERP experiment. Each participant of the cloze probability test received the first word of a given word pair, and was asked to provide the second word, so that the two words would make a meanigful expression. Due to the length of the test, all word pairs were divided into four blocks, so that each word was completed by 35 participants. If a given word pair was observed in the cloze probability test more than 3 times, the word pair was excluded from the final set and replaced with a new one. This procedure was repeated until the cloze probability for word pairs in all categories did not exceed 8%.

Meaningfulness In order to assess the meaningfulness of the stimuli, participants were asked to rate how meaningful a given word pair was on a scale from 1 (totally meaningless) to 7 (totally meaningful). The set of 228 word dyads was divided into four survey blocks in order to avoid the repetition of the target word within the same survey. Additionally, 76 meaningless word pairs were included in this normative study. The results revealed a main effect of utterance type, $[F(3, 387) = 1611.54, p < .001, \epsilon = .799, \eta_p^2 = .93]$. Pairwise comparisons revealed that literal word pairs were assessed as more meaningful ($M = 5.99, SE = .05$) than conventional metaphors ($M = 5.17, SE = .06$) ($p < .001$), and conventional metaphors were assessed as more meaningful than novel metaphors ($M = 4.09, SE = .08$)($p < .001$).

Familiarity Familiarity of each word pair was assessed in another normative study. Participants were asked to decide how often they had encountered the presented word pairs on a scale from 1 (very rarely) to 7 (very frequently). The set of 228 word dyads was divided into three survey blocks in order to avoid the repetition of the target word within the same survey. Again, a main effect of utterance type was found, $[F(2, 296) = 470.97, p < .001, \epsilon = .801, \eta_p^2 = .83]$. Pairwise comparisons showed that novel metaphors ($M = 2.15, SE = .07$) were rated as less familiar than conventional metaphors ($M = 2.97, SE = .08$), ($p < .001$), with literal expressions being most familiar ($M = 3.85, SE = .09$), ($p < .001$). Furthermore, conventional metaphors were less familiar than literal word dyads, ($p < .001$). It is crucial to note that although differences were observed between categories, all word pairs were relatively unfamiliar. This is visible in the mean score for literal word pairs, which are most familiar of all three categories, but at the same time relatively low in familiarity (below 4 on a scale where 6 and 7 represent very familiar items). The reason why familiarity was low in all three categories is the same as for the cloze probability test, i.e., that we intentionally excluded highly probable combinations.

Metaphoricity In order to assess the metaphoricity of the word pairs, participants were asked to decide how metaphoric a given word dyad was on a scale

from 1 (very literal) to 7 (very metaphoric). The set of 228 word dyads was again divided into three survey blocks in order to avoid the repetition of the target word within the same survey. The results revealed a main effect of utterance type, $[F(2, 198) = 588.82, p < .001, \epsilon = .738, \eta_p^2 = .86]$. Pairwise comparisons confirmed that novel metaphors (M = 5.00, SE = .06) were rated as more metaphoric than conventional metaphors (M = 3.98, SE = .06), ($p < .001$), and conventional metaphors were rated as more metaphoric than literal utterances $(M = 2.74, SE = .07), (p < .001)$.

3 Computational Modeling Method

In order to computationally model human judgment of the conceptual features of word dyads, we construct distributional semantic spaces where the proximity of word-vectors relates to their semantic similarity, and then explore the geometry of these spaces for ways of mapping relationships between words that are productive with regard to such conceptual, cognitive phenomena as metaphor. Specifically, we compare two different distributional semantic models to assess the difference in performance between a model that might be described as *static*, such as the one outlined in Section 3.1, versus one that is contextually *dynamic*, as is the intent with our own model as explained in Section 3.2.

For both our static and dynamic models, we train vectors on the English language version of Wikipedia. For the purpose of capturing word co-occurrences, we focus only on the descriptive content of Wikipedia pages, ignoring headers, lists, captions, and the like. Considering only sentences at least five words in length, we strip the corpus of punctuation, remove articles (*the*, *a*, and *and*), and remove parenthetical phrases, resulting in an overall corpus of approximately 7.5 million word types and 1.1 billion word tokens. For the construction of both models, we consider context windows of five words on either side of a target word, treating sentence endings as contextual boundaries as well. We take the 200,000 most frequently occurring words in the corpus as the vocabulary for both models, constructing one word-vector for each word in the vocabulary.

As our measure of semantic relatedness between two words, we take the cosine similarity between their corresponding word-vectors, in line with a number of other contemporary distributional semantic models [18, 23]. It should be noted that in the case of a normalised distributional semantic space, such as that described in Section 3.1, relations based on cosine similarity are equivalent to those based on Euclidean distance.

3.1 Word2Vec

As our primary point of comparison in this study, we use the Word2Vec distributional semantic model [22]. This model has achieved state of the art results on analogy completion tasks in particular, and has generally received widespread attention within the field of computational linguistics. A critical feature of the model is its deployment of a neural network to build a space of word-vectors. One result of this process is that the model's dimensionality cannot be interpreted:

Word2Vec treats a dimension as an arbitrary handle for pulling word-vectors into the desired relationship based on observations of co-occurrences in training data. Therefore, in comparison to our model described in Section 3.2, it is not possible to project dimensionally contextualized subspaces from a Word2Vec type model in a direct manner (while perhaps a separate neural network could be designed and trained specifically to perform this projection, this is beyond the scope of this paper).

Two different network architectures have been reported in the literature; here, we employ the *Skip-gram* architecture, consisting of a two-layer neural network which learns to predict context terms based on an input word, as this approach has been reported as performing particularly well on semantically oriented tasks [21]. The model takes the form of a set of word-vectors arrayed across the surface of a hypersphere. Here, we build a 300-dimensional space based on 10 passes over the corpus described above, with a negative sampling rate of 10. To assess the model's ability to capture the human metaphoricity judgment data, we then measure the cosine similarity between the word-vectors for each word in each word pair from the study described in Section 2.

3.2 Conceptual Discovery Model

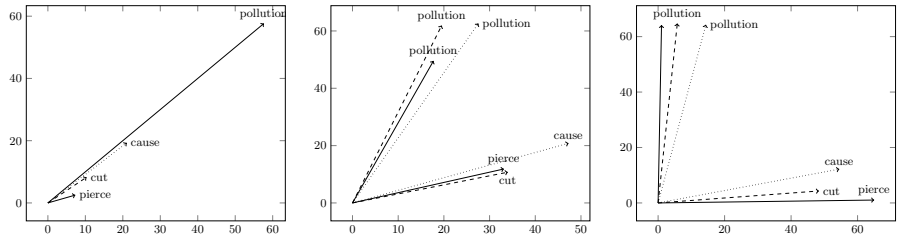
We compare the performance of the established Word2Vec semantic model with the output of a distributional semantic model which dynamically interprets input text to project context-sensitive subspaces from a sparse, high-dimensional base space. As described in detail elsewhere [1, 19], the Conceptual Discovery Model builds a base space populated by what might be described as literal statistical data about word co-occurrences as observed in a large-scale corpus: each dimension in the space corresponds directly to a co-occurrence term, and no matrix factorisation or other dimensional reduction technique is applied to this base space. Rather, each dimension c of a word-vector $\overrightarrow{w}$ is populated with a pointwise mutual information (PMI) score based on this equation, where $n_{w,c}$ represents the frequency at which word w is observed occurring within 5 words of word c, n_w is the independent frequency of w, n_c is the independent frequency of c, W is the total word count, and a is a smoothing constant:

$$\overrightarrow{w_c} = \log_2 \left(\frac{n_{w,c} \times W}{n_w \times (n_c + a)} + 1 \right) \tag{1}$$

We build a base space of roughly 7.5 million dimensions, corresponding to the number of word types in Wikipedia. From this base space, we dynamically pick 200-dimensional subspaces, specific to each word-pair in the study. Three different methodologies for projecting subspaces will be discussed below, but in each case, the input used to determine the projection is simply the pair of words involved in a potentially metaphoric dyad, and the projection is based on an analysis of the respective values of these inputs along any given dimension. The intuition behind this methodology is that subspaces consisting of dimensions which are mutually salient for both components of a dyad will capture something of the semantic context in which the candidate metaphor might be meaningful.

Within these subspaces, as with Word2Vec, we assess the relationship between the words in a word pair in terms of the cosine similarity between their two corresponding word-vectors. One of the primary considerations in the application of this model is therefore the method for selecting these subspaces.

Discovering conceptually relevant spaces We experimented with three different techniques for choosing subspaces from our base space, in each case focusing on the relationship between the target and source word in each dyad.



(a) Noun-Only Subspaces (b) Joint Subspaces (c) Independent Subspaces

Fig. 1. Here three different types of subspaces are presented, with three expressions involving the word "pollution" superimposed on each two-dimensional projection: the literal phrase "cause pollution", the conventional metaphor "cut pollution", and the novel metaphor "pierce pollution". Angles between the vectors for both words in each pair are measured for correlation with human judgments of the metaphoricity of each expression. Angles and vector lengths from the 200 dimensional subspaces we analysed are preserved in these projections. The word-vector for pollution is the same for all three versions of the noun-only space, since the other terms have no influence on the selection of dimensions here.

- **Noun-Only Subspaces**: the subspace is selected based only on associations with the target term: we take the 200 dimensions with the highest PMI value (as expressed in Equation 1) for the target (i.e., noun) in a given dyad.
- **Joint Subspaces**: selection is based on associations shared by the source and target terms: we select the 200 dimensions with the highest average PMI for both target and source terms in each dyad.
- **Independent Subspaces**: selection is based on independent associations with the source term and the target term, such that we select the 100 terms with the highest PMI values for the source term and the target term independently, and then merge these two sets of dimensions into a single 200 dimensional space.

An example of how a target and source dyads manifest in these three subspaces is shown in Fig. 1.

4 Results and Discussion

For both Word2Vec and CDM, cosine similarity values were computed for each word pair used in the behavioral study described above. Because the human ratings are the ground truth in this instance, Cosine Similarity is the dependent variable in each of the multiple regressions reported below. The three measures provided by human raters – Metaphoricity, Meaningfulness, and Familiarity – are the independent variables used in the analyses. The general aim is to identify which aspects of the word pairs (in terms of perceived metaphoricity, etc) are captured by cosine similarity in a given space. We also explore which type of subspace is best able to capture metaphor alone (that is, which space accounts for the most variability in human responses for metaphoricity).

We first report the results for Word2Vec, which are then used as a baseline against which to compare the results of our CDM model. The results for CDM are broken down by the type of underlying subspace.

4.1 Word2Vec results.

The results of the multiple regression analysis for Word2Vec indicated that the predictors accounted for a significant proportion of the variance in Cosine Similarity scores [$R^2 = .249$, $F(3, 224) = 24.81$, $p < .001$]. Metaphoricity significantly predicted Cosine Similarity scores, [$\beta = -0.25$, $t(224) = -3.09$, $p < .01$], as did Familiarity [$\beta = 0.22$, $t(224) = 2.65$, $p < .01$]. Low values of Metaphoricity tend to yield high values of Cosine Similarity, and low values of Familiarity tend to yield low values of Cosine Similarity. Meaningfulness was not a significant predictor in the regression.

First, these results confirm that Cosine Similarity does, in fact, capture more information than simply similarity about a given pair of terms: both Metaphoricity and Familiarity help to account for the variance in Cosine Similarity values for Word2Vec. Second, these results provide a standard by which we are able to compare the Conceptual Discovery Model's performance.

4.2 CDM results.

Unlike Word2Vec, the CDM model affords the discovery of different kinds of geometrically-defined subspaces. The crucial advantage of the CDM model is its ability to project a context-specific subspace geared towards capturing the semantics of situations in which a metaphor can be meaningfully applied. As such, our objective is to compare the performance of Cosine Similarity scores for detecting properties of metaphors using different techniques for constructing context-specific subspaces, in particular, the Noun-only, Joint, and Independent methods described in Section 3.2. The same multiple regression analysis as above was performed for these three CDM model configurations. Finally, the relationship between Cosine Similarity and Metaphoricity ratings is explored in more depth for the best performing model.

Noun-only subspaces and Joint subspaces The regression analysis for the Noun-only subspaces indicates that the predictors account for a limited proportion of the variance of Cosine Similarity scores, [$R^2 = .108$, $F(3, 224) = 9.04$, $p < .01$]. Metaphoricity significantly predicts Cosine Similarity scores, [$\beta = -0.30$, $t(224) = -3.48$, $p < .01$], where higher Metaphoricity ratings are associated with lower values of cosine similarity. The regression analysis for the Joint subspace does not yield any significant results, with $R^2 = .016$, $F(3, 224) = 1.19$, $p = n.s.$

Given the poor performance of the above models (which is quantified in the low R^2 for all four model configurations) compared with Word2Vec, we conclude that neither the Noun-only subspace nor the Joint subspace are suitable for capturing perceived metaphoricity of word pairs. Previous computational approaches to metaphor generation and interpretation [28, 31] have highlighted the fact that successful metaphors often result from situation in which the salient properties of one term (e.g., the target) are distinct from the salient properties of the other term. Therefore, the results may indicate that these two types of subspaces do not capture the necessary imbalance of salient properties between terms necessary to reflect metaphorical language. In other words, the Noun-only and Joint methods of delineating subspaces do not seem to select dimensions that are more salient for one term than the other, suggesting that Independent subspaces may provide a better way of capturing this information.

Independent subspaces The results of the multiple regression analysis for the Independently-constructed subspaces indicate that Metaphoricity accounts for a significant proportion of the variance in Cosine Similarity scores [$R^2 = .271$, $F(3, 224) = 27.73$, $p < .001$], and Metaphoricity significantly predicts Cosine Similarity [$\beta = -0.37$, $t(224) = -4.72$, $p < .001$].

Of the three multiple regressions reported here, this analysis accounts for the most variability in cosine similarity values, with an R^2 of .271, which is significantly higher than the regression for Word2Vec (where the R^2 was .249). Note that, interestingly, Familiarity is not significant in this analysis. The interpretation of this finding is discussed below in the General Discussion.

To visualize how the relationship between Cosine Similarity and Metaphoricity varies by utterance type (Conventional metaphor, Novel metaphor, and Literal word pair), a correlation analysis is shown in Fig. 2 for this best-performing model, with utterance types demarcated. Metaphoricity is inversely correlated with Cosine Similarity [$r = -.50, t = 32.49, p < .001$], such that word pairs rated as highly metaphorical tend to have low Cosine Similarity values. Novel metaphor word pairs, which participants rated highest for Metaphoricity, generally have low Cosine Similarity scores. This trend is shared by the Conventional metaphor word pairs, although the Metaphoricity scores tend to be slightly lower (this is confirmed by examining the averages of these two utterance types). Finally, Literal word pairs, which garnered the lowest ratings for Metaphoricity, tend to have slightly higher Cosine Similarity values overall.

In sum, whereas Cosine Similarity in Word2Vec is correlated with both Metaphoricity and Familiarity, the flexibility of our CDM model (specifically, the ability of our model to discover specific, conceptually-relevant spaces) allows

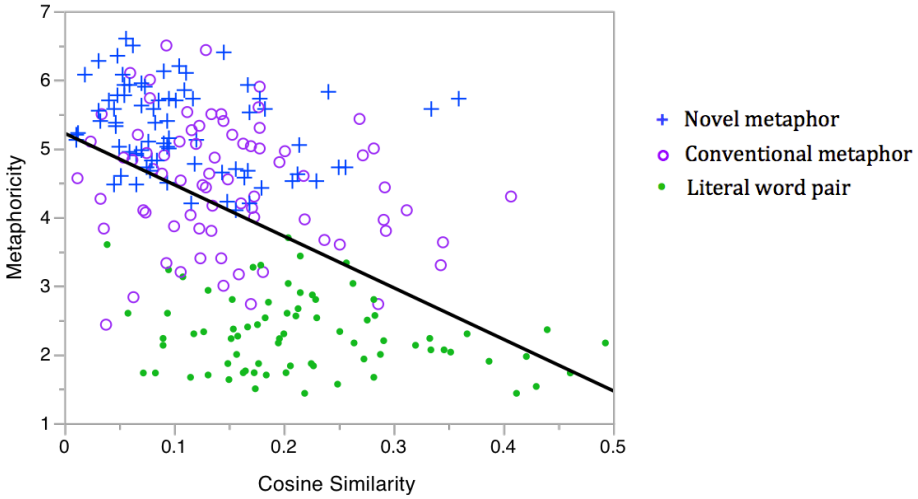


Fig. 2. Correlation between Cosine Similarity and Metaphoricity, including visualisation of Pair Types.

us to discover a space in which Cosine Similarity reflects only the *metaphorical* aspects of word pairs.

5 General Discussion

The set of results for Word2Vec and CDM offers important insight for the computational simulation of metaphoric language use. Firstly, for both Word2Vec and the Independent subspaces version of the CDM model, human ratings of Metaphoricity were able to account for a significant proportion of the variability in Cosine Similarity scores. Although only 24-27% of the variance was explained, it is important to consider that utterance type (which significantly influenced ratings) was *not* included in the statistical analyses above, because 1) the present research investigates the extent to which cosine similarity (alone) accounts for perceived metaphoricity between two terms, and 2) information about utterance type would not be available when applying our model in other contexts.

The performance of Word2Vec was used as a standard with which to compare the three variants of our CDM model. Both Familiarity and Metaphoricity were significant predictors of Cosine Similarity for Word2Vec, but for CDM, we were able to find a subspace that captures solely the perceived Metaphoricity of word pairs (because Metaphoricity was the only significant effect in the regression analysis). Furthermore, this Independent subspaces model performed best out of all of the models tested here, with an R^2 of .271 (while Word2Vec had an R^2 of only .249). Although this is only a modest improvement over Word2Vec, the difference does suggest that our CDM has a greater capacity to capture perceived metaphoricity. We therefore conclude that the Independent subspace

method offers both the most accurate and the most direct model of Metaphoricity (without confounding effects from Familiarity or Meaningfulness).

It is interesting to consider the comparative performance of the differently-configured CDM models, and explore why the Independent subspaces technique results in by far the best subspaces for mapping human judgments of metaphoricity to cosine similarity between word-vectors. In as much as a "property theoretic view of metaphor" [28, p. 56] has been laid out in computational terms, the expectation is that a successful computational model of metaphor will capture the way in which salient properties of a source are mapped to specific instances of a target. The subspaces generated by our model are intended to represent a conceptually relevant contextualisation of a lexical space: the dimensions of these subspaces consist of sets of words which taken independently offer only anecdotal glimpses into the way language is used, but which collectively can be understood as a certain *way of speaking* about a conceptually coherent topic.

So in a metaphorically relevant subspace, we hope to discover a *de facto* overlapping of some but not all of the properties of source and target. Rather than discover spaces where the mutual properties of two conceptual domains are already to some extent emphasised – as is the case with our Joint subspaces – we seek spaces where only a degree of overlap between the salient properties of each conceptual domain can be found, and where precisely this feature of a space is significant. This explains the efficacy of our Noun-only and, moreover, Independent subspaces in mapping human judgments of metaphor. In the case of the Noun-only subspace, we establish a context emphasizing the salient properties of the noun; to the extent that a verb expresses a conceptually paradigmatic action in this context, the cosine similarity between noun and verb word-vectors will be high, becoming lower as the noun-verb relationship becomes more metaphorical in nature. This phenomenon is considerably more evident in our Independent subspaces, where cosine similarity shows an inverse correspondence to the salient properties of each component of the dyad which have been merged into a single hybrid context.

It is worth noting that distributional semantic models have typically been applied to tasks involving the identification of word *similarity*, with the underlying intuition regarding these spaces being that similar words occur in similar contexts. Similarity here must be understood in a different light than the *familiarity* inherent in a word pairing: we might expect familiarity to correlate roughly with a tendency towards juxtaposition, and so a statistical measure of familiarity might emerge simply from calculating the PMI between two co-occurring words. Nonetheless, we must also note that words that tend to occur together will necessarily also tend to occur together *in the same context*, and so we might expect familiarity to emerge as a kind of artifact of this tendency in spaces geared towards similarity. It is therefore not surprising that a fairly standard distributional semantic model such as Word2Vec captures a degree of familiarity in measures of cosine similarity.

With this in mind, we might imagine a way forward towards building more nuanced subspaces particularly geared to prise apart judgments of metaphoricity. We could, for instance, investigate techniques for building subspaces that focus

primarily on the source component of a word pair – the verb, in the cases studied here – in order to draw out the salient properties of the source and then measure the degree to which these properties are typically transferable to a target. Finally, in future work we hope to use our findings regarding the geometric properties of subspaces to discover how people are likely to interpret new word pairs. In contrast to the research presented above, where human ratings were used to explain the variance in cosine similarity scores, this future direction will use cosine similarity scores for novel dyads to *predict* the degree to which human participants will perceive a given dyad as being metaphorical.

Acknowledgments

This research is supported by the project ConCreTe, which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733. This research has also been supported by EPSRC grant EP/L50483X/1.

References

1. Agres, K., McGregor, S., Purver, M., Wiggins, G.: Conceptualising creativity: From distributional semantics to conceptual spaces. In: Proceedings of the 6th International Conference on Computational Creativity. Park City, UT (2015)
2. Arzouan, Y., Goldstein, A., Faust, M.: Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. Brain Research 1160, 69–81 (2007)
3. Barnden, J.: Uncertainty and conflict handling in the ATT-Meta context-based system for metaphorical reasoning. In: Third International Conference on Modeling and Using Context. pp. 15–29 (2001)
4. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! In: ACL 2014 (2014)
5. Bowdle, B.F., Gentner, D.: The career of metaphor. Psychological Review 112(1), 193 (2005)
6. Clark, S.: Vector space models of lexical meaning. In: Lappin, S., Fox, C. (eds.) The Handbook of Contemporary Semantic Theory. Wiley-Blackwell (2015)
7. Feldman, J., Narayanan, S.: Embodied meaning in a neural theory of language. Brain and Language 84, 385–392 (2004)
8. Gibbs, R.W., Bogdanovich, J.M., Sykes, J.R., Barr, D.J.: Metaphor in idiom comprehension. Journal of Memory and Language 37(2), 141–154 (1997)
9. Goldstein, A., Arzouan, Y., Faust, M.: Killing a novel metaphor and reviving a dead one: ERP correlates of metaphor conventionalization. Brain and Language 123(2), 137–142 (2012)
10. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and Semantics, pp. 41–58. Academic Press, New York (1975)
11. Gutiérrez, E.D., Shutova, E., Marghetis, T., Bergen, B.K.: Literal and metaphorical senses in compositional distributional semantic models. In: Proceedings of the 54th Meeting of the Association for Computational Linguistics (2016, to appear)
12. Jankowiak, K., Naskręcki, R., Rataj, K.: Event-related potentials of bilingual figurative language processing. In: Poster presented at the 19th Conference of the European Society for Cognitive Psychology. Paphos, Cyprus (2015)

13. Kiela, D., Clark, S.: A systematic study of semantic vector space model parameters. In: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014. pp. 21–30. Gothenburg (2014)
14. Kiela, D., Hill, F., Clark, S.: Specializing word embeddings for similarity or relatedness. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2044–2048 (2015)
15. Kintsch, W., Bowles, A.R.: Metaphor comprehension: What makes ametaphor difficult to understand? Metaphor and Symbol 17(4), 249–262 (2002)
16. Kutas, M., Federmeier, K.D.: Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (ERP). Annual Review of Psychology 62, 621 (2011)
17. Lai, V.T., Curran, T.: ERP evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors. Brain and Language 127(3), 484–496 (2013)
18. Levy, O., Goldberg, Y.: Linguistic regularities in sparse and explicit word representations. In: 18th Conf. on Computational Natural Language Learning (2014)
19. McGregor, S., Agres, K., Purver, M., Wiggins, G.: From distributional semantics to conceptual spaces: A novel computational method for concept creation. Journal of Artificial General Intelligence (2015)
20. McGregor, S., Purver, M., Wiggins, G.: Words, concepts, and the geometry of analogy. In: Workshop on Semantic Spaces at the Intersection of NLP, Physics and Cognitive Science (2016)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of ICLR Workshop (2013)
22. Mikolov, T., tau Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 246–251 (2013)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Conf. on Empirical Methods in Natural Language Processing (2014)
24. Schütze, H.: Dimensions of meaning. In: Proceedings of the 1992 ACM/IEEE conference on Supercomputing. pp. 787–796 (1992)
25. Shutova, E., Teufel, S., Korhonen, A.: Statistical metaphor processing. Computational Linguistics 39(2), 301–353 (2012)
26. Turney, P.D., Patel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141–188 (2010)
27. Utsumi, A.: Computational exploration of metaphor comprehension processes using a semantic space model. Cognitive Science 35(2), 251–296 (2011), http://dx.doi.org/10.1111/j.1551-6709.2010.01144.x
28. Veale, T.: A service-oriented architecture for metaphor processing. In: Proceedings of the Second Workshop on Metaphor in NLP. pp. 52–60 (2014)
29. Wolff, P., Gentner, D.: Evidence for role-neutral initial processing of metaphors. Jnl. Experimental Psychology: Learning, Memory, and Cognition 26(2), 529 (2000)
30. Wolff, P., Gentner, D.: Structure-mapping in metaphor comprehension. Cognitive Science 35(8), 1456–1488 (2011)
31. Xiao, P., Alnajjar, K., Granroth-Wilding, M., Agres, K., Toivonen, H.: Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In: Proceedings of the 7th International Conference on Computational Creativity (ICCC). Paris, France (2016)