

A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

A thesis to be submitted to the University of London for the degree
of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary, University of London
United Kingdom

September 2017

Abstract

This thesis describes a novel methodology, grounded in the distributional semantic paradigm, for building context sensitive models of word meaning, affording an empirical exploration of the relationship between words and concepts. Anchored in theoretical linguistic insight regarding the contextually specified nature of lexical semantics, the work presented here explores a range of techniques for the selection of subspaces of word co-occurrence dimensions based on a statistical analysis of input terms as observed within large-scale textual corpora. The relationships between word-vectors that emerge in the projected subspaces can be analysed in terms of a mapping between their geometric features and their semantic properties. The power of this modelling technique is its ability to generate ad hoc semantic relationships in response to an extemporaneous linguistic or conceptual situation.

The product of this approach is a generalisable computational linguistic methodology, capable of taking input in various forms, including word groupings and sentential context, and dynamically generating output from a broad base model of word co-occurrence data. To demonstrate the versatility of the method, this thesis will present competitive empirical results on a range of established natural language tasks including word similarity and relatedness, metaphor and metonymy detection, and analogy completion. A range of techniques will be applied in order to explore the ways in which different aspects of projected geometries can be mapped to different semantic relationships, allowing for the discovery of a range of lexical and conceptual properties for any given input and providing a basis for an empirical exploration of distinctions between the semantic phenomena under analysis. The case made here is that the flexibility of these models and their ability to extend output to evaluations of unattested linguistic relationships constitutes the groundwork for a method for the extrapolation of dynamic conceptual relationships from large-scale textual corpora.

This method is presented as a complement and a counterpoint to established distributional methods for generating lexically productive word-vectors. Where contemporary vector space models of distributional semantics have almost universally involved either the factorisation of co-occurrence matrices or the incremental learning of abstract representations using neural networks, the approach described in this thesis preserves the connection between the individual dimensions of word-vectors and statistics pertaining to observations in a textual corpus. The hypothesis tested here is that the maintenance of actual, interpretable information about underlying linguistic data allows for the contextual selection of non-normalised subspaces with more nuanced geometric features. In addition to presenting competitive results for various computational linguistic targets, the thesis will suggest that the transparency of its representations indicates scope for the application of this model to various real-world problems where an interpretable relationship between data and output is highly desirable. This, finally, demonstrates a way towards the productive application of the theory and philosophy of language to compu-

tational linguistic practice.

Glossary

base space A high dimensional, sparse vector space of word-vectors, delineated in terms of dimensions of co-occurrence statistics.

context The situation – environmental, cognitive, perceptual, linguistic, and otherwise – in which an agent finds itself and applies language to meaning.

contextual input A set of words characteristic of a conceptual category or semantic relationship used to generate a subspace for the modelling of semantic phenomena.

dimension selection The process of contextually choosing a subset of dimensions in order to project a subspace from a base space.

co-occurrence The observation of one word in proximity to another in a corpus.

co-occurrence statistic A measure of the tendency for one word to be observed in proximity to another across a corpus.

co-occurrence window The boundary defining the proximity within which two words are considered to be co-occurring, typically a distance in terms of words within a sentence.

methodology The process of building base spaces from observations of co-occurrences within a corpus and contextually projecting subspaces through dimension selection.

model An application of methodology to a particular linguistic task or experiment, sometimes including task specific statistical analysis techniques.

subspace A context specific lower-dimensional projection from a base space, effectively mapping semantic relationships to a context by way of the geometric relationships between word-vectors.

word-vector A high-dimensional geometrically situated semantic representation of a word, constructed as an array of co-occurrence statistics.

Table of Contents

Abstract	i
Glossary	iii
Table of Contents	iv
List of Figures	v
List of Tables	vi
3 Metaphor and Coercion	1
3.1 An Experiment on Metaphor	2
3.1.1 Methodology and Results	4
3.1.2 The Geometry of Metaphor	9
3.1.3 Generalising the Model	13
3.2 An Experiment on Coercion	15
3.2.1 Methodology and Results	16
3.2.2 The Geometry of Coercion	19
3.2.3 Adding Sentential Context	20
3.3 Interpretation and Composition in Context	20
References	22

List of Figures

3.1	Receiver Operating Characterisation for Metaphor Classification	8
3.2	Metaphors In Space	12
3.3	Receiver Operating Characterisation for Coercion Classification	19
3.4	Three dimensional projections of word-vectors and generic vectors in sub-spaces for pairs at the extents and in the middle of the literal-metaphorical spectrum.	20

List of Tables

3-A	Context Sensitive and Static Model F-Scores for Metaphor Classification .	5
3-B	F-Scores for Metaphor Classification of Unseen Adjectives	7
3-C	Comparative Metaphor Classification Statistics	9
3-D	Top Independent Features for Metaphor Classification	10
3-E	Most Predictive Feature Vectors for Metaphor Classification	11
3-F	Scoring Metaphoricity Based On Classification Data	14
3-G	Context Sensitive and Static Model F-Scores for Coercion Classification .	17
3-H	F-Scores for Coercion Classification Testing on Unseen Verbs	18
3-I	Top Independent Features for Coercion Classification	19
3-J	Comparison of the seven most effective features for coercion classifica- tion in 2x2 word, 400 dimensional subspaces for INDY versus VERB based dimension selection.	20

Chapter 3

Metaphor and Coercion

In this chapter, I will extend the empirical work on exploring the application of my context sensitive distributional semantic models to two semantic phenomena which involve the application of words in situations where their meanings are in some sense conceptually altered: *metaphor* and *semantic type coercion*. The precise definitions of these terms, which are not without nuance, was explored in Chapter ?? and will be reintroduced in subsequent sections. As an overview, the distinguishing characteristic of these phenomena is that they involve cases where what might be thought of as the stable, encyclopedic understanding of some word sense – a *dictionary definition* of a word, so to speak – is in some way appropriated or subverted in order to, among other things, transfer information via the attributional conduits connecting figurative source to literal target.

My hypothesis is that, because figurative language always involves the contextual specification of word meaning, context sensitive geometries of lexical representations should provide an appropriate framework for identifying when this type of semantic phenomenon is in effect. Fraser (1993) demonstrates empirically that metaphor interpretation is, when a metaphor is presented to a subject out of context, an ambiguous exercise, and, to the extent that interpretations of de-contextualised metaphors can be predicted, the predicting factors are themselves culturally relative. Along similar lines, Bouveret and Sweetser (2009) propose that metaphor production involves the contextual alignment of overlapping semantic frames, and that this alignment likewise imports structure associated with one frame into the domain of another, evident in, for instance, the additional transposition of syntactic constraints from source to target. From a cognitive perspective, this coordinates a contextual theory of metaphor with the work on conceptual frames from Barsalou (1992,9) discussed at the end of the previous chapter in the context of judgements of semantic similarity. From a modelling perspective, this suggests that a methodology for projecting semantic spaces where context specific perspectives can reveal *ad hoc* perspectives on semantic relationships should be a productive approach to identifying figurative language.

The idea that metaphor and metonymy are both instances of “a connection between two things where one term is substituted for another,” (Gibbs Jr., 1993, p. 260) will quickly call to mind the premise of distributional semantics: if the motivation for building

vector space models of word co-occurrence statistics is that related words have similar co-occurrence tendencies, then figurative language might be construed as a special case in which unrelated or at least conceptually divergent words are likewise found in similar sentential situations. The question, then, is whether statistical characteristics of the particular co-occurrences profiles selected by words with different meanings are predictive of figurativeness. A naive hypothesis might be that word combinations that are figurative should simply be further apart in a semantic space than word combination that are literal. If related words have similar co-occurrence profiles, then maybe unrelated words, for instance words with different conceptual entailments, should have less similar co-occurrence profiles. This conjecture, however, is belied first of all by the fact that, in the type of corpus containing a broad range of examples of language use necessary for building distributional semantic models, figurative language will already be built into the data (and at the end of this chapter I will argue, in line with, for instance, Gibbs (1994), that figurative language is going to be built into any sample of language no matter how small or basic). A second problem is that, specifically to overcome the problems with modelling semantic relationships merely in terms of collocations, distributional semantics compares the co-occurrence profiles of words rather than their direct relationships, and it seems likely that word combinations prone to metaphoric interpretation might very well have at least overlapping profiles.

So the objective of the experiments reported in this chapter will be to explore the ways in which and the degrees to which a more fleshed out statistical description of contextually selected distributional semantic subspaces can reveal figurative language. As with the experiments on relatedness and similarity reported in the previous chapter, in addition to the relationship between target word-vectors in the subspaces they select, the statistical properties of the selected dimensions themselves will also be examined. And, again as with previous results, the instrument of analysis will be the geometric features of the subspaces in question, with, again, particular attention paid to the way in which the sets of features can collectively indicate figurative language. The two primary datasets explored represent binary decisions about metaphoricity and coercion respectively, and so my models will be applied to classification tasks here. In the case of metaphor, I test whether a model learned based on classification data is generalisable to graduated human ratings of metaphoricity. With the coercion data, I will examine whether the addition of information about sentential context enhances the classification of word pairs. I will conclude the chapter with a reflection on some of the theoretical implications of the strongly positive results described here.

3.1 An Experiment on Metaphor

As pointed out by Shutova et al. (2012), statistical approaches to metaphor identification and interpretation have generally been formulated in the context of the *conceptual metaphor* theory of ?. This model is founded on the principle that “we systematically use inference patterns from one conceptual domain to reason about another conceptual domain,” (ibid, p. 246). Metaphors are then the mechanism for performing the mapping between these domains, and as such cut right to the core of cognitive processes. Statisti-

cal models of metaphor have accordingly treated metaphors as transformations of lexical representations, and vector space models of distributional semantics have naturally leant themselves to this type of approach. The construction of representations with the potential to interact with one another in semantically productive ways has in turn lent itself to the development of models that consider the compositional nature of metaphor, effectively treating the metaphor itself as a transformation of the underlying representations. So Utsumi (2011) constructs candidate metaphor-vectors by calculating the centroid of a number of vectors derived from an analysis of a noun-vector and a predicate-vector learned through latent semantic analysis, and then uses the spatial relationships between these composed vectors to analyse the metaphoricity of certain phrases. Hovy et al. (2013) similarly consider composition in their approach to metaphor classification, in this case by combining word-vector type representations with a model trained to identify metaphor based on dependency trees of sentences labelled for metaphoricity.

In the tradition of work on compositional distributional semantics explored by the likes of Mitchell and Lapata (2010), Baroni and Zamparelli (2010), and Coecke et al. (2011), among others, semantic types such as adjectives and verbs are modelled as tensors which perform transformations on nouns, which are modelled as vectors. In the normal run of things, compositional models therefore represent, for instance, noun phrases modified by adjectives as the product $A\vec{n}$, where A is a matrix representing an adjective learned from observations of attested instances of the adjective with other noun word-vectors. So the phrase *black dog* becomes a word-vector in the same space as the representation of just *dog*, and can be compared quantitatively and geometrically with other phrases such as *white dog* or *big cat* and so forth. In the case of metaphor, these transformations are expected to map the word-vector representing metaphoric phrases into a region corresponding to the semantic domain of the original noun-vector modified by a metaphoric interpretation of the word associated with the tensor of a modifier or a predicate. So, for instance, in a model that effectively captures metaphoricity, the composition of the vector space representations corresponding to *brilliant light* would map to a region of space where comparisons between phrases like *dark illumination* and *red glow* are productive, while *brilliant child* might be expected to map into the proximity of *stupid boy* and *boring girl*.¹

The data that I will use in this section to test my methodology was originally presented by Gutiérrez et al. (2016), along with an accompanying experiment on a novel model. It consists of 8,592 adjective-noun pairs, spanning 23 adjectives chosen for their membership in six different broad semantic categories that are prone to both literal and metaphoric use: so, for instance, *bitter*, *sour*, and *sweet* are considered constituents of the category TASTE. There are 3,473 different noun types used, with only 141 types, represented by 640 tokens, occurring in both literal and metaphoric phrases. Each pair has been rated as either literal or metaphoric by a pair of human annotators, with inter-annotator agreement measuring at Cohen's $\kappa = 0.80$; 4,593 of the pairs have been judged metaphorical. This dataset was conceived as something of an expansion of the similar but smaller corpus of adjective-noun phrases annotated with binary metaphoricity classifications presented by Tsvetkov et al. (2014) (and those authors tested their own data

¹It should be noted that such a methodology at this point begins to assume dim shades of Gärdenfors's (2000) conceptual spaces, with different compositions inherently defining different regions of the space.

with an assortment of models, achieving highest f-scores by applying a random forest classifier to the features of an existing library of distributional semantic word-vectors).

In their own experimental treatment of the data, Gutiérrez et al. constructed a pair of compositional models in the mode of Baroni and Zamparelli (2010), learning adjective matrixes A to map from noun-vectors to noun-adjective phrase-vectors extracted from observations of co-occurrences of both nouns and phrases in a corpus. By creating separate tensor representations for literal and metaphoric instances of a given adjective, the authors can then compare the relationships between the vectors resulting from a noun-vector composed with literal and metaphoric senses of an adjective-vector to try to determine whether a given phrase would generally be classified as a metaphor or a literal expression by comparing the respective compared vectors to the phrase-vector as observed in the corpus. In a further attempt to generalise the method, and, notably, to apply the conceptual metaphor theory of Lakoff and Johnson (1980) to their computational model, the authors learn matrices performing linear transformations from literal to metaphoric adjective-noun compositions and then compare the similarity between observed phrase-vectors and literal composed vectors versus transformed literal composed vectors to determine whether a given phrase is metaphoric or not.

The data described by Gutiérrez et al. will serve as the basis for testing my own context sensitive distributional semantic methodology's ability to classify phrases as literal or metaphoric, and the results of this experiment will be described in the following section. My hypothesis is that metaphor, and indeed all figurative language, is fundamentally entangled with the context mutually indicated by the representations of the words participating in the composition being analysed. In fact, I think that part of what is captured by the model described by Gutiérrez et al., and indeed a number of other researchers investigating statistical methods for metaphor classification, is precisely that there is a context inherent in the linear algebraic dynamics of composable lexical representations, and this is something which many researchers explicitly recognise. But I also think that the explicit projection of context specific semantic subspaces, the mainstay of my methodology, should provide an ideal testing ground to discover the way in which statistical geometry can directly broadcast the presence or absence and even potentially the degree of metaphor inherent in a given phrase. The following sections will test this hypothesis using a similar methodology to that applied to semantic relatedness and similarity in the previous chapter.

3.1.1 Methodology and Results

My own methodology is clearly less committed to maintaining distinct representations for different semantic types than the compositional models described above, instead modelling all words as untagged word-vectors based on their co-occurrences as observed across a large scale corpus. This feature of my research is in part theoretically motivated: in line with Langacker (1991), and *contra* the grammatic nativism or exceptionalism that has been a mainstay in theoretical linguistics (?), I would like to investigate the possibility that "grammar is fully and appropriately describable using only symbolic units, each having both semantic and phonological import," (ibid, p. 290). In other words, the syntactic

<i>window dimensions</i>	2x2				5x5			
	20	50	200	400	20	50	200	400
JOINT	0.839	0.860	0.878	0.881	0.840	0.862	0.880	0.886
INDY	0.821	0.839	0.855	0.860	0.817	0.840	0.858	0.867
ZIPPED	0.839	0.864	0.876	0.878	0.833	0.854	0.873	0.880
ADJECTIVE	0.771	0.860	0.828	0.845	0.781	0.804	0.828	0.837
NOUN	0.819	0.861	0.843	0.847	0.806	0.821	0.838	0.843
SVD	0.685	0.703	0.703	0.697	0.677	0.694	0.687	0.684
SG	0.679	0.676	0.679	0.673	0.664	0.665	0.672	0.656
CBOW	0.669	0.681	0.677	0.672	0.669	0.673	0.677	0.671

Table 3-A: F-scores for metaphor identification based on a stratified ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

component of a natural language might be described in terms of the entanglements of the meaning-making structures – the lexical semantic representations – that arise in the course of language use, or maybe even as emergent properties of these entanglements.

With this in mind, I will approach the problem of metaphor classification with a similarly statistical and geometric methodology as was applied to relatedness and similarity in the previous chapter, outside of any *prima facie* model of syntax or compositionality. For every pair of words in the data produced by Gutiérrez et al. (2016), I generate subspaces of 20, 50, 200, and 400 dimensions using the JOINT, INDY, and ZIPPED techniques, projected from 2x2 and 5x5 word co-occurrence window base spaces. This data specifies a distance role for each word, one being a metaphoric source (the adjective) and the other being a target (the noun): so, for instance, a *bitter loss* is a loss, but presumably not one with an actual taste, and so the noun *loss* co-opts something of the quality of bitterness into its own conceptual domain. As such, it might be useful to generate subspaces based simply on an analysis of the word-vectors corresponding to the adjective and the noun respectively. I do this by simply selecting the top d dimensions, in line with the dimensionality parameter for each model, for the term in question, and these spaces are labelled ADJECTIVE and NOUN in the results that follow.

In each subspace, I extrapolate the same 34 geometric features described in Table 3-J and applied in the previous chapter in the semantic relatedness and similarity experiments. Again because of the semantic asymmetry of the relationship between the input terms, an additional seven features are also available in these spaces: the adjective-vector norm divided by the noun-vector norm (A/B), likewise the lengths of the vectors between the adjective and the generic points divided by the lengths for the noun-generic-point vectors ($\overline{AC}/\overline{BC}$, $\overline{AM}/\overline{BM}$, and $\overline{AX}/\overline{BX}$), and the corresponding fractions of the normalised versions of these points ($\overline{A'C'}/\overline{B'C'}$, $\overline{A'M'}/\overline{B'M'}$, and $\overline{A'X'}/\overline{B'X'}$). These additional measures might offer a sense of whether there are statistical tendencies that are specific to the semantic role being played by a word moving from literal to metaphorical relationships, and we might expect this to be particularly evident in the spaces selected by either the noun or the adjective on their own. As with the subspaces of relatedness and similarity, I normalise each feature across all word pairs to have means of 0 and standard deviations of 1.

In order to test the capacity of the geometric features of my subspaces to identify metaphor, I perform a stratified ten-fold cross-validated logistic regression taking these features as independent variables and learning to predict the classifications assigned to the word pairs in the dataset. Balanced f-scores based on the precision and recall of my various dimensional selection techniques as well as static SVD factorisations of my base spaces and the `word2vec` models are reported in Table 3-A. The first thing to note is the strong performance across the board of the context sensitive methodology: the model based on my strongest performing subspace (JOINT, 5x5 window, 400 dimensions) substantially outperform the strongest versions of the static models (the SVD 5x5, 400 dimension model) with $p < .005$ based on a permutation test. The context sensitive models perform better, but only marginally better, in the 5x5 word window subspaces, suggesting that most of the useful information about the semantic properties that indicate a metaphoric projection are captured by the profile of terms co-occurring in close proximity to the target words. That this trend is reversed for the static spaces, with 2x2 word window spaces doing a bit better, further indicates that the peripheral information of wider ranging co-occurrences is specifically useful for a context sensitive analysis.

The JOINT technique gives the strongest results, suggesting that subspaces delineated in terms of co-occurrence dimensions mutually salient to both input terms offer the best platform for analysing metaphoricity. This makes sense: in the case of metaphor versus literalness, it is the co-occurrences that both words have in common that position their respective word-vectors in an indicative relationship relative to one another and the subspace overall. So for instance the co-occurrences salient to both *sweet* and *fruit* will have a particular conceptual profile that will not be evident in the dimensions jointly selected by *sweet* and *revenge*; this effect will be less evident for dimensions independently salient to each word. ZIPPED subspaces, where there will be at least some information about both words along every dimension, accordingly score almost as well as JOINT subspaces, with the INDY subspaces falling further behind.

Interestingly, the ADJECTIVE and NOUN spaces classify metaphor most accurately in 50 dimensional subspaces projected from the 2x2 word window base space. To the extent that part-of-speech can be a component of the analysis of these models, we can expect the smaller co-occurrence window to produce statistics that are more indicative of a particular grammatical class. The degradation of classification at higher dimensionalities for the smaller co-occurrence window setting is a little surprising, and it's worth noting that the INDY subspaces, which are basically blends of the ADJECTIVE and NOUN subspaces, don't exhibit the same tendency. In this case, it would seem the whole really is greater than the sum of the parts, with the dimensional selection of one word providing at least a degree of useful information about the other word not available in spaces salient to a single term. A similar pattern emerges for the static spaces: the SVD, SG, and CBOW models all produce the most accurate classifications in 2x2 word window, 50 dimensional subspaces. One way to explain this is that more ambiguous information about word use begins to leak in at higher dimensionalities, serving to obscure the more standard indications available in either the most salient dimensions or the dimensions containing the most information about variance across the corpus.

There is another possibility to consider regarding the adjectives in this dataset in par-

<i>window dimensions</i>	2x2				5x5			
	20	50	200	400	20	50	200	400
JOINT	0.815	0.837	0.854	0.855	0.816	0.837	0.858	0.863
INDY	0.778	0.793	0.828	0.835	0.774	0.805	0.829	0.842
ZIPPED	0.810	0.838	0.847	0.854	0.799	0.828	0.844	0.853
ADJECTIVE	0.606	0.709	0.750	0.777	0.698	0.697	0.757	0.707
NOUN	0.806	0.808	0.828	0.833	0.796	0.812	0.824	0.829
SVD	0.679	0.691	0.695	0.690	0.665	0.674	0.678	0.676
SG	0.668	0.664	0.659	0.657	0.659	0.656	0.644	0.638
CBOW	0.657	0.665	0.665	0.661	0.656	0.660	0.666	0.660

Table 3-B: F-scores for metaphor identification with each of the conceptual categories identified by Gutiérrez et al. (2016) treated as a separate fold for cross-validation.

ticular: as there are only 23 different adjective types, each adjective is observed multiple times in both metaphoric and non-metaphoric contexts. It is therefore possible that, in any given fold of the cross-validation of a classifier, the model might be learning how to guess whether a specific adjective is involved in a metaphor rather than something more general about the statistical geometry of metaphoricity. In order to avoid this trap, I reorganise the data into tranches based on the adjective in each pair, I use the eight conceptual categories outlined by Gutiérrez et al. (2016) in order to structure this new partitioning.² I use each of these eight new sets of word pairs as a fold in a cross-validated logistic regression, such that the adjective in each phrase in each test set has not been observed in the training data.

Table 3-B presents the results from this reshuffled version of the experiment. The f-scores for metaphor classification returned by the context sensitive models are down slightly, but the difference is only marginally significant

XXX SIGNIFICANCE

The major change here is, as expected, in the ADJECTIVE subspaces: clearly when only information from the adjective in each word-pair is used to train a model, prior observations of a specific word type in the context of some other composition is a benefit. There is also a minor decrease in performance for the static models, which is interesting in that it indicates that, even when a single distance metric is used to classify metaphoricity, observations of a word in training help to subsequently test phrases involving that word. It is worth noting that of the 8,584 noun tokens spread across 3,473 noun types, 1,588 types, represented by 6,724 tokens, occur in more than one of the tranches delineating the conceptual categorisations of the adjectives, so it is possible that there is a small extent of learning to classify phrases based on previous observations of specific nouns.

In order to take a closer look at the way that different techniques model this data, and in line with the metaphor classification work of Tsvetkov et al. (2014), Figure 3.1

²Gutiérrez et al. (2016), identifying a similar problem, likewise develop a second model that learns metaphors as mappings between domains rather than just from noun-vectors to phrase, though their methodology requires them to use a reduced version of the data.

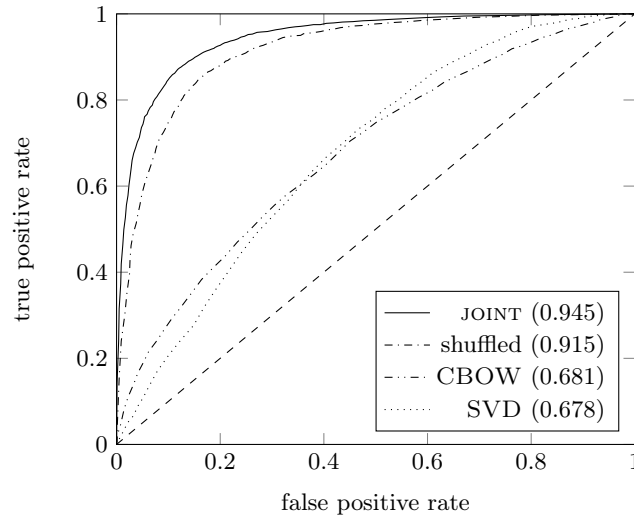


Figure 3.1: Receiver operating characteristic plots for a selection of models, with the area under the curve for each model type indicated in the legend.

illustrates receiver operating characteristic curves for four versions of the approaches that have been described here: the JOINT technique with 400 dimensional, 5x5 word subspaces, the same technique applied to the version of the data shuffled to avoid training and testing on the same adjectives, and the CBOW and SVD models for the optimally performing 50 dimensional, 2x2 word window subspaces. True positive versus false positive rates are correlated at 99 increments in terms of the value of the output of a logistic regression model at which a phrase is determined to be metaphoric. The outcomes visualised here tell a similar story to Tables 3-A and 3-B, with the area under the curve statistics indicating a strong distinction between the context sensitive techniques and the static models. Perhaps the most interesting thing to note is the overall smoothness of the curves, which suggests a steady relationship between precision and recall at various classification thresholds.

With the trade off between true and false positives in mind, Table 3-C presents precision, recall, f-score, accuracy, and Cohen’s kappa scores for the same models plotted in Figure 3.1. The trend to notice here is that context sensitive and static models tend to favour recall over precision (and the slight preference for precision in the JOINT 400 dimensional, 5x5 word subspaces for the shuffled version of the data reported here is an anomaly, as other approaches to that data exhibit the tendency towards higher recall). This evident enthusiasm for classifying phrases as metaphoric is a reflection of the data itself, which is slightly skewed towards metaphoric phrases, as described above and indicated in the performance of the majority class baseline, and this is reinforced by the relatively low accuracy scores for both context sensitive and static non-compositional distributional semantic models. It is noteworthy, then, that the model described by Gutiérrez et al. (2016) actually scores better for precision than recall, suggesting it actually tends to under-predict metaphoricity. This could perhaps be expected as a general distinction between statistical models based on unannotated data such as mine, which will arguably tend to favour a majority class, versus likewise statistical models operating

	<i>precision</i>	<i>recall</i>	<i>f-score</i>	<i>accuracy</i>	<i>kappa</i>
JOINT	0.879	0.894	0.886	0.877	0.753
shuffled	0.873	0.865	0.862	0.854	0.678
SVD	0.631	0.794	0.703	0.641	0.265
CBOW	0.638	0.721	0.677	0.632	0.253
Gutiérrez et al. (2016)	0.842	0.793	0.817	0.809	0.618
baseline	0.535	1.000	0.697	0.535	0.000

Table 3-C: Full classification statistics results for the models tested here as well as the results from the original literature and the majority class (metaphor) baseline.

on theoretically motivated mappings between representations, which have an apparent propensity for zeroing in with confidence on the properties of a compositional transformation that are indicative of metaphor—but at the expense of sometimes missing what might be considered outliers. In the same spirit, the jumpier nature of the receiver operating characteristic plots presented by Tsvetkov et al. (2014) is quite possibly an artefact of the decision points inherent in heuristically mapping model features from human made knowledge bases.

As a final point of comparison with other approaches to metaphor classification, I will return briefly to the unannotated character of my lexical representations. One of the most powerful features of the methodology described here is its ability to build a somewhat general model of a semantic phenomenon from a sufficiently comprehensive dataset, and the strong Cohen’s kappa score of the best performing subspace selection technique, which begins to approach the aforementioned inter-annotator agreement level of $\kappa = 0.80$, is a testament to this. Following an analysis of the specific geometry of metaphor in the next section, Section 3.1.3 will assess the ability of my methodology to generalise even further from this data to a broader range of metaphors and to moreover move from classification to gradation based on observations of merely binary judgements of metaphoricity. For now, I simply note that it is remarkable that data about nothing more than the way that words tend to be collocated can, with the aid of a mechanism for contextualisation, reveal so much about the nature of the semantic relationship between the lexical components of an previously unseen phrase.

3.1.2 The Geometry of Metaphor

In this section, I will explore the geometric features which prove most productive in the classification of metaphor. As with relatedness and similarity in the previous chapter, I begin by examining the capacity of independent features to predict metaphor. Rather than a proper logistic regression involving multiple independent variables fed into a non-linear function, this analysis amounts to choosing a cut-off point in terms of the value of each feature separating literal and metaphoric phrases in the subspaces which an analysis of their corresponding word-vectors delineate. So the f-scores reported in Table 3-D can be understood as indicating the degree to which the values of a given geometric feature separate the dataset into distinct categories corresponding to human judgements

JOINT		INDY		ZIPPED	
$\mu(A, B)$	0.787	C	0.767	$\mu(A, B)$	0.788
C	0.771	C/M	0.749	C	0.771
$\mu(A, B)/M$	0.764	$\angle AMB$	0.747	$\mu(A, B)/M$	0.769
$\angle COX$	0.762	C/X	0.746	X	0.767
X	0.762	$\mu(A, B)$	0.734	$\mu(A, B)/X$	0.759
ADJECTIVE		NOUN			
$\mu(A, B)/M$	0.745	$\mu(A, B)$	0.756		
$\overline{AC} : \overline{BC}$	0.736	C	0.747		
$\overline{AC}/\overline{BC}$	0.734	$\mu(A, B)/X$	0.728		
$\mu(A, B)/X$	0.732	$\mu(A, B)/M$	0.721		
$\angle ACB$	0.730	C/X	0.721		

Table 3-D: Independent f-scores from the metaphor classification data for top five features of each subspace type for 5x5 word co-occurrence window, 400 dimension subspaces.

of metaphoricity.

The scores themselves reflect the trend observed in Table 3-A and 3-B: the JOINT and ZIPPED subspaces produce features that are particularly good at classifying metaphor, with a decrease in performance in the INDY subspaces and then another step down in the single-word subspaces. None of the scores themselves come close to the levels of discrimination achieved by the models learned from full feature vectors, though

XXX SIGNIFICANCE

In terms of the actual features indicated by this analysis, two in particular figure prominently in one way or another, namely, the mean of the word-vector norms $\mu(A, B)$ and the norm of the central-vector C . In the first instance, the role of the relationship between word-vectors and the origin of the spaces that their salient co-occurrence dimensions delineate is once again reflective of the preliminary findings on conceptual geometry described in Chapter ??, where norm was seen to be an effective mechanism for defining a region of conceptual constituency. In the case of the distance of the central vector from the origin, the emergence of this feature, as well of the appearance of the norms M and X as components of various strongly predictive tendencies, indicate that here, as with similarity in the previous chapter, characteristics of dimensions outside of the situation of any particular word-vector along them might be in themselves indicative of metaphor: some words might simply be more likely to co-occur in the context of metaphoric language, and co-occurrence statistics should provide a handle for examining this tendency.

To further delve into the statistical geometry of metaphor, and in line with the results on relatedness and similarity described in the previous chapter, I once again search the state space of possible combinations of features to find the optimal feature vector for classifying metaphor in context sensitive subspaces. This is again treated as a beam search problem, though the search space expanded at each level of the search tree is here limited to the top 500 combinations of features given the larger size of the data being modelled. Table 3-E presents the optimal seven feature combinations discovered for the

		10-fold ($f = 0.869$)	shuffled ($f = 0.830$)
DISTANCES			
word-vectors	-		-
generic vectors	$M = -1.448$		-
ANGLES			
word-vectors	$\angle ACB = -0.775$		-
normalised	-		-
generic vectors	$\angle COX = -1.618$	$-0.271 = \angle COM$	
	$\angle COM = 0.974$	$0.045 = \angle MOX$	
MEANS			
word-vectors	$\mu(\overline{AM}, \overline{BM}) = -1.124$	$-1.007 = \mu(\overline{AC}, \overline{BC})$	
normalised	-		-
RATIOS			
word-vectors	-	$0.492 = \overline{AM} : \overline{BM}$	
		$-0.620 = \overline{AX} : \overline{BX}$	
normalised	-	$-0.168 = \overline{A'C'} : \overline{B'C'}$	
FRACTIONS			
word-vectors	$\overline{AC}/\overline{BC} = 0.325$		-
generic vectors	$M/X = 1.305$	$0.252 = A/B$	

Table 3-E: The seven most predictive features for metaphor classification, compared between ten-fold and sight-unseen cross-validation of logistic regression on statistics extrapolated from 5x5 word window, 400 dimensional JOINT subspaces.

5x5 word window, 400 dimensional JOINT subspaces based on both a standard ten-fold cross-validation and the version of the data shuffled in order to test on data not observed in each training phase. The f-scores achieved by these combinations of features, reported next to the respective labels at the top of the table, indicate a marginal decrease in the overall performance as compared to the full featured models of subspaces, but the results are still strong.

Angles between generic vectors, which were already evident as independently predictive features in Table 3-D, have a strong effect here, with the strong negative correlation of $\angle COX$ in the ten-fold cross-validation in particular suggesting that maximal values tend to be relatively similar across dimensions jointly selected by literal adjective-noun combinations, pulling the line of X closer to the centroid described by C . To put this differently, as pairs become more metaphoric, they tend to also become less consistent in the type of dimension that they co-select, as evidenced in the increasing variance in the maximum values of these dimensions. Perhaps the most interesting thing to observe here, though, is the strong correlation between ratios of word-vector to generic vector distances in the case of the version of the data shuffled to test on unseen adjectives, but not in the case of the stratified cross-validation. The positive correlation with the balance of the distances from the word-vectors to the mean vector M means that subspaces where the word-vectors have a relatively even relationship to the weighted centre are, in fact, more metaphoric (and their relationship to the maximum vector is comparatively less balanced, with this vector in turn being less central to the space per the observations regarding

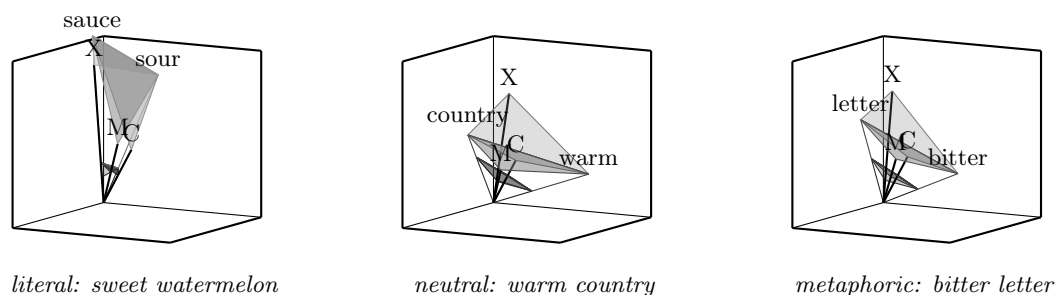


Figure 3.2: Three dimensional projections of word-vectors and generic vectors in subspaces for pairs at the extents and in the middle of the literal-metaphorical spectrum, taken from 5x5 word window, 400 dimensional subspaces selected using the JOINT technique.

$\angle COX$). But more generally, it is noteworthy that the balance between word vectors and generic vectors is informative about metaphoricity specifically in models tested on unseen adjectives: this balance is in effect a projection into space of quotients of joint probabilities of observing words and co-occurrence terms divided by the typical or maximal probabilities of being observed with the co-occurrence terms, and from it we can infer that these quotients are generally predictive of metaphor in context, even without word-specific training data.

Figure 3.2 presents visualisations by way of three dimensional projections of word-vectors and generic vectors from 400 dimensional JOINT subspaces selected from the 5x5 word window base space.³ In the example of the uncontroversially literal phrase *sweet watermelon*, the word-vectors are characteristically far from the origin and close to one another, corresponding to the predictivity of $\mu(A, B)$ in particular. At the other extent of the spectrum, the highly metaphoric phrase *bitter letter* is characterised by a dropping of the word-vectors and a widening of the angle between them; the generic vectors, meanwhile, are now further from the origin relative to the word-vectors that select the subspace and, at the same time, draw closer to one another in particular at the normalised layer of the subspace. But most interestingly, at a relatively neutral point, occupied by the intriguingly ambiguous phrase *warm country*, which the logistic regression trained on these subspaces assigned a score of close to 0.5, there is actually evidence of an intermediary widening out of the overall array of points even as the word-vectors remain fairly far from the origin.

The exciting thing about this last observation is that it suggests that, rather than existing on a linear or even monotonic scale, metaphor may itself actually be a multi-dimensional phenomenon, with a characteristic particular to highly ambiguous word combinations that is to some extent separate from the statistical features of straightforward literalness and clear cut metaphoricity. The broad arrangement of word-vectors in space engendered by the contextualisation of the phrase *warm country*, in contrast to

³These projections have been rendered using the same regression technique as applied to the images for related word pairs in the previous section, but the coordinates of X have been divided by 1.5 instead of 2.

the relatively tight relationship of the generic vectors, can be interpreted as revealing an uncertainty regarding the semantic properties being transferred in this small composition, corresponding to a drifting of the word-vectors and a contracting of the generic vectors across the jointly selected co-occurrence profile. Here, once again, the statistical geometry of a subspace can be productively mapped to a theoretical statement about the nature of a semantic phenomenon as characterised by a selectively contextual and quantitative representation of observations about the way that words are used, by and large outside of any strong preconditions symbolically encoded in the computational framework.

3.1.3 Generalising the Model

One of the interesting things about feature-based classification is that there is typically an inherent commitment to degree of class membership, even when the training data used to build a model is simply binary. This is true of any model which uses, for instance, a logistic regression technique for determining class, as there is a cut-off point along the spectrum of model output and a corresponding proximity to that point for any given sample, and it is especially obvious when the features of the model are actually geometrical measures. In this section, I will apply the models learned from the the Gutiérrez et al. (2016) data to another dataset designed to assess metaphor as a matter of degree rather than simply as a binary situation, and a dataset that additionally deals with a different type of metaphor in terms of composition. The question explored here is whether the geometric features of context specific distributional semantic analysis of word-vectors will provide binary classification models with adequate information for projecting metaphoricity along a continuous scale.

The data used for this experiment was originally reported by ?, and was used to train a model based on an earlier version of methodology as described by ?. This data consists of 228 predicate-object word pairs selected to cover three degrees of metaphor, consisting of literal pairs such as *announce willingness*, conventionally metaphoric pairs such as *cut pollution*, and novel metaphors such as *smell excuses*. 102 human participants provided metaphoricity scores on a seven point Likert scale, and the average scores were compiled into the dataset that will be used to test models learned from the geometric features output by my context sensitive methodology.⁴ Specifically, I will experiment with two different classification model techniques. In the first instance, I will take the output of the logistic regression described above, trained on the Gutiérrez et al. (2016) data, as assigning probabilities to the metaphoricity of an input word pair, and I will in turn measure the degree to which these probabilities correlate with the degree of metaphoricity collectively assigned by human raters. In the second instance, I'll use the binary metaphor classification data to train a support vector machine.⁵ Applying a radial basis function kernel, I analyse the correlation between distance from the discriminatory hyperplane and the human ratings. In both cases, and in line with results reported in the previous chapter, Spearman's correlations are the unit of analysis.

⁴Studies were also conducted to gather ratings for *familiarity* and *meaningfulness*, but those ratings will not be modelled in this thesis.

⁵This is implemented using the python scikit-learn SVC module.

<i>features</i>	1	3	5	7	9	full
<i>logistic regression</i>						
JOINT	0.368	0.355	0.033	0.085	0.279	-0.033
ADJECTIVE	-0.377	0.355	0.044	0.513	0.511	0.335
<i>support vector machine</i>						
JOINT	0.352	0.359	0.042	0.045	0.243	0.158
ADJECTIVE	-0.170	0.247	-0.021	0.407	0.418	0.236

Table 3-F: Spearman’s correlation with human verb-noun metaphoricity scales judgments based on logistic regression and support vector machine models trained on adjective-noun classification data, taking feature vectors of various lengths as independent variables.

Table 3-F presents results for both modelling techniques, focussing on features extrapolated from 5x5 word window, 400 dimensional subspaces using the JOINT approach and through an analysis on the adjective in each word-pair from the input data. Feature vectors of various lengths, picking the optimal geometric features for each dimensional selection technique, are used to feed input to each model. In terms of the models trained on features from JOINT subspaces, there is a clear trend towards strong performance with one or three features, weaker performance with five or seven features, stronger performance again with nine features, and then a drop-off again in the full featured space. The relatively low performance with the full set of features is not particularly surprising: there is clearly an encroaching incidence of generalisation error here as the models become flooded with data about various and certainly collinear statistical features of contextual geometry. At the shallow end of the feature selection parameters, on the other hand, the single measure $\mu(A, B)$ (per Table 3-D) once again points to the efficacy of word-vector norm as a predictive characteristic of contextualised co-occurrence subspaces.

The really remarkable outcome here, though, is the very strong performance of the models learned from the top seven and nine features extracted from subspaces selected by PMI values of the adjective word-vectors alone. This is particularly interesting given that the data being tested actually consists of a different type of grammatical relationship, namely, predicate-object pairs. It would seem, then, that the co-occurrence dimensions most salient to either verbs or adjectives generate a geometry in which their relationship to potential arguments can play out in similar ways in terms of the metaphoricity inherent in the semantic context: the interaction between the selecting vector, the noun-vector, and the generic vectors translates from one type of composition to another in an isomorphic way. This explanation, including the claim that the mapping of predictive features from one type of metaphor to the other is to a large extent isomorphic, is supported by the particularly strong performance of the logistic regression at seven and nine dimensions, where the logistic function takes a polynomial with coefficients learned in the training phase as direct input. The more complex non-linearity afforded by the support vector machine appears to actually somewhat confound the mapping from verb-noun to adjective-noun phrases—though the difference between the correlations at nine dimensions is not statistically significant at $p = .104$ based on a Fisher r-to-z transformation.

The one area where a support vector machine provides a clear improvement in perfor-

mance is in the full dimensional models extrapolated from JOINT subspaces. In this case, it would seem that the radial basis function classification actually does a better job of avoiding the overfitting in a higher dimensional feature space. But, putting questions of model choice aside, there is clear evidence here for the generality of the contextual geometry of metaphor, and also a strong case for the appropriateness of machine learning techniques for providing an appropriate mechanism for the computational manipulation of co-occurrence information to build a more nuanced model of degree of metaphor based on relatively rudimentary classification data. Crucially, it is the context sensitivity of my methodology that facilitates the exploration of a multi-dimensional feature space in which the non-linear nuances of this particular semantic phenomenon can be discovered; a model providing a singular static relationship between lexical representations could not offer the context specific underpinning for generating a geometry replete with interpretable statistical features. Finally, there are signs here to invite further research, and indeed some grounds for hoping that a context sensitive approach might have the scope for handling more sophisticated tasks such as metaphor interpretation and generation.

3.2 An Experiment on Coercion

In this section, I will apply my methodology to the classification of a phenomenon closely related to metaphor, namely, *semantic type coercion*, by which the semantic type of a noun is reassigned in the course of a verb taking that noun as an argument. So, for instance, in phrases like *denied wrongdoing* or *heard footsteps*, the nouns in play are standing in for a conceptually relevant but different type of noun, and the literal versions of these phrases would go something like *denied committing wrongdoing* or *heard the sound of footsteps*, where the verbs select arguments of types along the lines of ACTIVITY and PERCEPTION respectively. This phenomenon is often referred to as *logical metonymy*, identifying it as a subspecies of the more general figurative phenomenon metonymy by which a thing is denoted by a conceptually related lexical representation.

Coercion is one of the semantic phenomena targeted by Pustejovsky's (1995) theory of a *generative lexicon*, by which nouns are semantically modelled as having a *qualia structure* which maps out the way that a thing relates to itself, the world, and the agents interacting with it in that world on four different levels of abstraction, with the general objective of arriving at "a model of meaning in language that captures the means by which words can assume a potentially infinite number of senses in context, while limiting the number of senses actually stored in the lexicon," (ibid, p. 104). In terms of coercion, qualia provide the basis for a process of *projection* by which a variety of semantic types can be extracted from a complex type (or a *dot object* in Pustejovsky's lingo) in order to fulfil the typing requirements of a predicate in open ended ways. The model that emerges here – one built on dynamically interactive lexical semantic representations contingent on some sort of general conceptual context – begins to look like the general linguistic stance that has motivated my own methodology.

This theoretical commitment suggests a schematic by which a symbol manipulating system might begin to get a handle on productive and context sensitive lexical repre-

sentations of things in the world. To this end, Jezek and Hanks (2010) have described an ontology based on a computational analysis of co-occurrence patterns designed to facilitate the modelling of what is ultimately a sliding scale of statistically enhanced semantic representations, or “shimmering lexical sets,” (ibid, p. 19), as the authors put it. Applying a similar notion that coercion is probabilistic rather than discreet, Lapata and Lascarides (2003) use co-occurrence statistics to try to predict the verbs which, in the role of for instance participles, successfully resolve instances of coercion. And, under the rubric of *logical metonymy*, Shutova et al. (2013) expand upon the work of Lapata and Lascarides by extracting verb senses from WordNet to build a class based model, to some extent recapitulating the categorical distinctions that characterise many theoretical approaches to coercion. The motivation behind this last system is the apt observation that, in the case of coercion, “humans are capable of interpreting these phrases using their world knowledge and contextual information,” (Shutova et al., 2013, 11:2).

Returning to the theoretical issues regarding grammaticality raised earlier in this chapter, the analysis of coercion within the framework of the generative lexicon points to something more like a graduated typology, sliding from specific instances of processes, things, and the like to more general conceptual categories and finally to entire classes of words. As Langacker (1991) has pointed out, there is a lurking ambiguity in grammatical class distinctions, with various conceptual schema existing in any natural language for moving between classes: so, to borrow an example from Langacker, phonological and symbolic dynamics facilitate a conceptually coherent progression from *sharp* to *sharpen* to *sharpener*, and the rules that are extrapolated as an explanatory framework for such transitions are just a way of systematising the cognitive networks that underpin these linguistic

6

With this in mind, my hypothesis is that, as with metaphor in the previous section, a syntactically neutral statistical model with a context generating capacity should be able to capture

3.2.1 Methodology and Results

The data which will be used to test my methodology in this section was originally presented by ? as a task for the ongoing International Workshop on Semantic Evaluation series of computational semantic modelling challenges. The data consists of 2,071 sentences each containing a marked verb and object, with the object classified as either coercive or not. The verbs cover various conjugations of five different verb stems, each identified as selecting for a different semantic type as an argument: the verbs (and the semantic type selected) are *arrive* (LOCATION), *cancel* (EVENT), *deny* (PROPOSITION), *finish* (EVENT), and *hear* (SOUND). The objective, then, is to train a model to indicate that the phrase *finish the party* is not coercive, in as much as we accept that *party* denotes a member of the conceptual category EVENT, whereas *finish the food* is because what is actually being finished is the event of eating food, not the food itself. For the purposes of

⁶One is also reminded of ?’s (?) quip regarding “grammatical fictions,” (ibid, ¶307).

<i>window dimensions</i>	2x2				5x5			
	20	50	200	400	20	50	200	400
JOINT	0.604	0.619	0.630	0.657	0.634	0.672	0.673	0.691
INDY	0.666	0.677	0.703	0.693	0.652	0.660	0.707	0.679
ZIPPED	0.568	0.624	0.610	0.647	0.596	0.625	0.658	0.663
VERB	0.664	0.675	0.698	0.704	0.631	0.652	0.699	0.700
NOUN	0.601	0.628	0.643	0.633	0.518	0.565	0.603	0.641
SVD	0.511	0.523	0.539	0.412	0.521	0.409	0.483	0.563
CBoW	0.498	0.508	0.531	0.493	0.496	0.544	0.535	0.496
SG	0.518	0.565	0.575	0.529	0.534	0.523	0.583	0.557

Table 3-G: F-scores for coercion identification based on a ten-fold cross-validated logistic regression taking geometric features of various subspace types as input.

the original presentation the data is split into a training set and a testing set of roughly equal size, but questions of the most meaningful partitioning of the data will be discussed below.

Two amendments are made to the data as presented. First, of the 2,071 verb-object pairs, 197 contain multi-word objects not compatible with the vocabulary used for my model, reducing the total number of word pairs to 1,874, 543 of which are considered coercive. Second, of these remaining computable word pairs, 903 are duplicates (they are presented in unique sentences, but for the first phase of analysis here only verb-noun pairs will be consider; sentential context will be addressed below). This leaves a total of 971 word pairs, 376 of which are deemed coercive. As with the metaphor data in the previous section, I train a logistic regression model to discriminate between regular argument selection and coercion. I once again take the two words being analysed as input to generate a number of different context specific distributional semantic subspaces, treating the 34 geometric features outlined in Table 3-J plus the seven additional fractional features specific to asymmetric input terms described above in Section 3.1.1 as the independent variables of the regression analysis.

Table 3-G presents the f-scores derived from the precision and recall results of a ten-fold cross-validation of these logistic regression models. Most obviously, these numbers are considerably lower than the comparable results for metaphor outlined in Table 3-A, but this is to some extent mitigated by the relative scarcity of instances of coercion in the data: a minority class baseline always classifying word pairs as coercive would, based on the above data statistics, give $f = 0.558$. The top score of $f = 0.708$ for the context sensitive models, achieved by the 5x5 word window, 400 dimensional verb-only dimensional selection technique, is significantly better than the baseline with

XXX significance

Of the three dimensional selection techniques that use both words as input, the INDY method does best here (as opposed to the JOINT technique for metaphor), indicating that coercion plays out most clearly in the relative co-occurrence profiles of word-vectors across dimensions salient to words individually rather than jointly, and this is corroborated by

<i>window dimensions</i>	2x2				5x5			
	20	50	200	400	20	50	200	400
JOINT	0.338	0.397	0.362	0.381	0.345	0.428	0.404	0.386
INDY	0.454	0.386	0.436	0.459	0.369	0.350	0.411	0.410
ZIPPED	0.256	0.297	0.363	0.358	0.324	0.352	0.377	0.357
VERB	0.233	0.334	0.361	0.448	0.307	0.401	0.352	0.336
NOUN	0.306	0.398	0.406	0.401	0.243	0.293	0.317	0.340
SVD	0.295	0.252	0.276	0.126	0.217	0.173	0.301	0.288
CBoW	0.368	0.329	0.248	0.162	0.302	0.316	0.245	0.177
SG	0.349	0.333	0.281	0.194	0.366	0.351	0.316	0.229

Table 3-H: F-scores for coercion identification taking each verb stem type as a separate fold of a cross-validation.

the strong performance of the VERB and NOUN subspaces generated based on an analysis of either one of the two input words.

In line with the metaphor results is the poor performance of the static models, which generally do somewhat worse than the baseline and substantially worse than the context sensitive models. Of particular note is the decline of the SVD models and the comparative ascent of the **word2vec** skip-gram methodology: the sentential context predicting mechanism of the skip-gram approach seems to better capture the typological relationships between predicates and arguments than a principal component analysis of the dimensional variance in a base space of co-occurrence statistics. But in fact, the results here are across the board less regular in their relationship to parameters of dimensionality and co-occurrence window size, with a more even distribution of relatively high and low scores for both 2x2 and 5x5 word co-occurrence window models, and comparatively strong outcomes occasionally popping up for 20 or 50 dimensional spaces. The seemingly erratic output of the model gives an overall impression of an unanchoring between the statistics of co-occurrence and the semantic phenomenon being explored here. Perhaps in the case of coercion, or at least in terms of the data sampled here, many predicate-object combinations are, regardless of the influence of the verb on the noun’s conceptual situation, too conventional for type shifts to be detected in a meaningful way in terms of co-occurrence profiles.

Another telling feature of these results is the quite strong performance of the subspaces selected by an analysis of the verbs alone. In fact, this is likely to be an artefact of the data itself: only five different verb stems are used, and some are arguably marked by their own semantic peculiarities, with, for instance, *finish* coercing 152 out of the 252 arguments it takes in the data, where the rate for *deny* is only 29 out of 183 instances. In order to find out if the models being tested here are actually just learning, in one way or another, specific rules about particular inputs, I rearrange the data into five folds corresponding to the five verb types present, training a model on each combination of four different verbs and then testing the model on the classifications of word-pairs involving the fifth. F-scores are reported in Table 3-H.

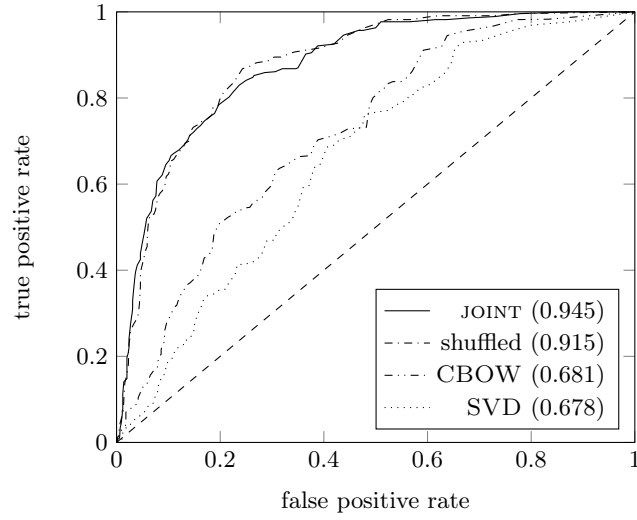


Figure 3.3: Receiver operating characteristic plots for a selection of models for coercion classification, with the area under the curve for each model type indicated in the legend.

JOINT		INDY		ZIPPED	
A/B	0.544	A/B	0.447	$\angle AXB$	0.450
$\mu(\overline{A'C'}, \overline{B'C'})$	0.543	$\angle AOB$	0.442	$\mu(A, B)/C$	0.312
$\angle AXB$	0.537	$\mu(\overline{A'M'}, \overline{B'M'})$	0.435	$\mu(\overline{A'C'}, \overline{B'C'})$	0.276
$\mu(\overline{A'X'}, \overline{B'X'})$	0.527	$\mu(\overline{A'C'}, \overline{B'C'})$	0.417	$\angle AOB$	0.261
$\mu(\overline{A'M'}, \overline{B'M'})$	0.514	$\mu(\overline{A'X'}, \overline{B'X'})$	0.385	$\mu(\overline{A'X'}, \overline{B'X'})$	0.249
		VERB		NOUN	
		$\mu(\overline{A'M'}, \overline{B'M'})$	0.315	$A : B$	0.531
		$\overline{AC}/\overline{BC}$	0.314	A/B	0.531
		A/B	0.252	$\mu(\overline{AB})/C$	0.484
		$\mu(\overline{A'C'}, \overline{B'C'})$	0.248	$\angle AMB$	0.420
		$\angle A'M'B'$	0.238	$\angle ACB$	0.375

Table 3-I: Independent f-scores from the coercion classification data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces, validated on unobserved verbs.

3.2.2 The Geometry of Coercion

('0.683', [(('mAMB', '0.848'), ('fABM', '-0.825'), ('aCOM', '0.407'), ('mAXB', '-0.349'), ('aAOB', '-1.756'), ('nACB', '0.537'), ('eACB', '0.207'))])

		INDY ($f = 0.676$)	VERB ($f = 0.687$)
DISTANCES			
word-vectors	-		-
generic vectors	-		-
ANGLES			
word-vectors	$\angle AMB = -0.831$		-
normalised	-		-
generic	$\angle COM = 0.197$		-
	$\angle COX = -0.159$		-
MEANS			
word-vectors	$\mu(A, B) = 0.919$		$-0.817 = \mu(A, B)$
normalised	-		-
RATIOS			
word-vectors	-		-
normalised	$\mu(\overline{A'C'} : \overline{B'C'}) = 0.015$		
FRACTIONS			
word-vectors	-		$0.490 = \mu(\overline{AB})/M$
			$0.443 = A/B$
			$-0.046 = \mu(\overline{AB}/C)$
normalised	-		$-1.863 = \overline{A'C'} : \overline{B'C'}$
			$-1.602 = \overline{A'M'} : \overline{B'M'}$
generic vectors	$C/M = -0.838$		$0.974 = M/X$
	$M/X = 0.234$		

Table 3-J: Comparison of the seven most effective features for coercion classification in 2x2 word, 400 dimensional subspaces for INDY versus VERB based dimension selection.

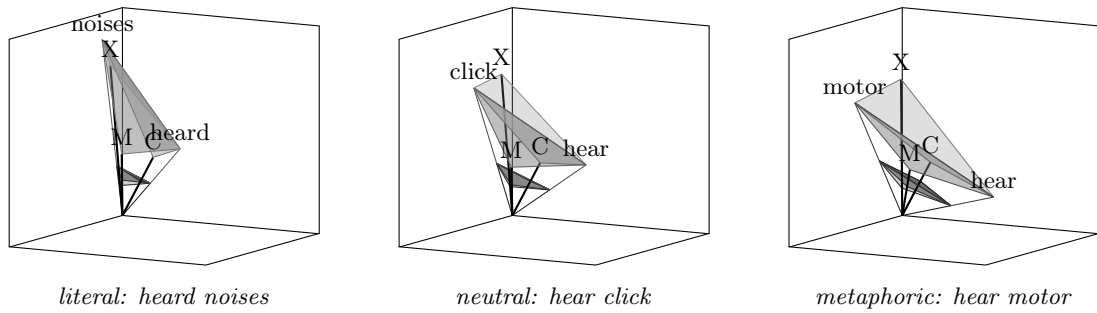


Figure 3.4: Three dimensional projections of word-vectors and generic vectors in subspaces for pairs at the extents and in the middle of the literal-metaphorical spectrum.

3.2.3 Adding Sentential Context

3.3 Interpretation and Composition in Context

One of the tricky things about figurative language is its ephemerality: if we stare at it for long enough through a theoretical lens, it seems to vanish, as is evident in the deflationary

case made by ?. But on the other hand, if we ask someone in street whether the phrase *buy a story* is more metaphoric than *buy a book*, we can reasonably expect the answer will almost always be “yes”, and it would be a mistake to dismiss the evidence that in a colloquial sense some compositions are clearly metaphoric, and others are clearly not. This raises a challenging point with regard to the comparison between metaphor and coercion, the two instances of figurative language explored in this chapter: is metaphor perhaps to some extent a more overt case of coercion, or maybe a specific case that is in some way or another a little more subtle? Part of the problem here is that the distinctions between these phenomena begin to exceed the capacity for what can reliably be quantified about language in a clinical setting, with evaluative criteria that will depend on the opinion of an expert which comes pre-packaged with inevitable biases.

In fact, it is tempting to go so far as to say that figurative language is identified precisely as those instances of language where recourse to a conceptual context is necessary to interpret a lexical composition, and furthermore that the degree of figurativeness correlates with the extent of context construction involved in an interpretation. This proposition is in line with Shutova’s (2015) empirical work treating metaphor interpretation as a mechanism for classification

This, then, raises a valid question: is the role of figurative language exclusively, or even for that matter primarily, to port attributes from one conceptual domain to another? Or is what metaphor does, as ? has famously suggested, really about something more fundamentally phenomenological than just the efficient transmission of propositions? So, where, for instance, ? sees polysemy as an intermediate stage bridging the progress from literal to metaphoric usage, my methodology leaves itself open to the possibility that all usage is, in fact, first and foremost pragmatic, and only secondarily lexicalised. By this interpretation, words have semantic affordances in terms of their potential to convey cognitive content intersubjectively, and they are picked up and used in much the same way that a cognitive agent might adapt an object designed or just perceived as being for one purpose as an implement in another activity—using a shoe as a hammer, for example, or a chair to fend off a lion. The cognitive foregrounding of this nascent theory can be found in the ecological psychology of ? and ?, and the linguistic correlary seems to be in line with what psycholinguists inspired by biosemiotics such as ? are saying about the way that language is primarily about affording cognitive value to interlocutors, including but hardly limited to truth values.

This theoretical speculation is a potential extrapolation of my methodology rather than a precondition for it, and is offered primarily as an example of how this statistical approach might become a component of productive line of philosophical enquiry. The point, though, is that with a geometric methodology, relationships between lexical semantic representations can be recast as Gibsonian affordances: there is a mechanism for the direct perception of opportunities for meaning making in the actual layout of the statistical environment

References

- Baroni, M. and Zamparelli, R. (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.
- Barsalou, L. W. (1992). Frames, concepts, and conceptual fields. In Lehrer, A. and Kittay, E. F., editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 21–74. Lawrence Erlbaum Associates, Hillsdale, N.J.
- Barsalou, L. W. (1993). *Theories of Memory*, chapter Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols. Lawrence Erlbaum Associates, Hove.
- Bouveret, M. and Sweetser, E. (2009). Multi-frame semantics, metaphoric extensions and grammar. *Annual Meeting of the Berkeley Linguistics Society*, 35(1):49–59.
- Coecke, B., Sadrzadeh, M., and Clark, S. (2011). Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis*, 36(1-4):345–384.
- Fraser, B. (1993). *Interpretation of novel metaphors*, pages 307–341. Cambridge University Press, 2 edition.
- Gärdenfors, P. (2000). *Conceptual Space: The Geometry of Thought*. The MIT Press, Cambridge, MA.
- Gibbs, Jr., R. W. (1994). *The Poetics of Mind*. Cambridge University Press.
- Gibbs Jr., R. W. (1993). *Process and products in making sense of tropes*, page 252–276. Cambridge University Press, 2 edition.
- Gutiérrez, E. D., Shutova, E., Marghetis, T., and Bergen, B. K. (2016). Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Hovy, D., Srivastava, S., Kumar, S., Sachan, J. M., Goyal, K., Li, H., Sanders, W., and Hovy, E. (2013). Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.
- Jezek, E. and Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4:7–22.
- Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Langacker, R. (1991). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Mouton de Gruyter, Berlin.
- Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.
- Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.
- Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

- Shutova, E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.
- Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):11:1–11:28.
- Shutova, E., Teufel, S., and Korhonen, A. (2012). Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.
- Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *ACL (1)*, pages 248–258. The Association for Computer Linguistics.
- Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2):251–296.