[[Something Pithy]]:
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

July 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 betweendata 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

Chapter 3

Relatedness and Similarity

In Chapter **??**, I laid out the theoretical groundwork for statistical context sensitive models of lexical semantics, and in Chapter **??** I described the actual methodology for building such models, accompanied by a preliminary proof of concept involving conceptual entailment. In this chapter, I will present the first set of experiments designed to evaluate the utility of this methodology. These experiments are intended to probe the productivity of a context sensitive, geometric approach to building a computational model of lexical semantics based on statistics about word co-occurrences. Beyond testing my models' performances on some well-travelled datasets, this will provide an opportunity to explore whether different components of the methodology and, moreover, different aspects of geometric output lend themselves to modelling related but distinct semantic phenomena.

So, moving into familiar computational linguistic territory, I will explore my methodology's performance on two different phenomena: *relatedness* and *similarity.* Each of these objectives have provided reliable but distinct evaluative criteria for computational models of lexical semantics. One of the hypotheses I will put forward regarding my methodology is that the geometrically replete subspaces generated by my contextualisation techniques should provide features for the simultaneous representation of related, diverse, and sometimes antagonistic aspects of language. Experimenting with these established datasets will provide a platform for exploring the ways in which different features of a semantic structure projected into one of my contextualised subspaces shift as the relationships inherent in the generation of the subspace likewise change, and this will in turn lead to some searching questions about the importance of context in the computational modelling of these particular semantic phenomena in the first place.

A fundamental objective for a general semantic model is a mechanism for measuring the relatedness inherent in semantic representations. The distributional hypothesis itself is framed in terms of the relatedness between words: if words that tend to have a similar co-occurrence profile should also tend to have similar meaning, then, in some sense of the word, *similarity* is what is being captured by the word-vectors that populate a distributional semantic model. There is, however, an ambiguity at play in terms of what exactly it means for two words to denote things that are semantically *related*, and when this designation should include the more specific category of *similarity* (or, for that matter, other types of relatedness such as *meronymy*, *analogy*, even *antonymy*, and so forth). So, for instance, the words *tiger*, *claw*, *stripe*, *ferocious*, and *pounce* are all clearly related in the way that they trace out aspects of a very specific conceptual space of TIGERNESS, but none of them are similar in the way that *tiger*, *lion*, and *bear* are all commensurable constituents of a space of WILD ANIMALS.

The compilation of data for the purpose of testing the ability of computational models to identify semantic relationships between words has tended to focus on the general case of relatedness rather than more nuanced similarity, if sometimes simply through a failure to specify between the two. The methodology for generating this data goes something like this: human participants are given a set of pairs of words and asked to quantify, for instance, the "similarity of meaning" (**?**, p. 628) in each pair, or "how strongly these words are related in meaning," (**?**, p. 124). **?** use both the terms *similarity* and *relatedness* in the instructions for generating their WordSim353 data, analysed below, ultimately asking evaluators to rank words from being "totally unrelated" to "very related";[1] **?** used only the term *relatedness* in their instructions, with no mention of *similarity*.[2] **?** have discussed the uncertainty inherent in human ratings produced in this manner, pointing out that judgements of similarity and relatedness can be subjective and task specific.

Relatively recently, researchers have made a concerted effort to generate data that focusses on word similarity specifically, rather than a less clearly defined notion of relatedness. **?** have taken the widely used WordSim data and split it into two overlapping sets of word pairs, one intended to reflect a range of judgements on word similarity and the other judgements on relatedness, based on human evaluations of the types of relationships inherent in each word pair. Subsequently **?** have created Their SimLex999 dataset by extracting word pairs from an existing set of word associations, sampling from a range of conceptual relationships, and then giving human evaluators detailed instructions casting similarity in terms of degree of synonymity. These datasets have proven more resistant to highly accurate modelling through standard distributional semantic approaches—indeed, an interesting corollary to the distinction between relatedness and similarity has been the development of *knowledge based* versus *corpus based* techniques for modelling these semantic phenomena (see **??**, for a discussion), with corpus based, or statistical, techniques proving more suited to modelling relatedness rather than similarity.

My thoroughly statistical methodologies will be initially tested on the WordSim353 data in order to explore my subspaces' capacities for capturing semantic relatedness and the SimLex data in order to explore how it handles similarity. The models learned based on this data will then be applied to alternate datasets for relatedness and similarity to gauge their generality. The most valuable outcome of this set of experiments, however, will be the comparison between the models learned for each of these related but distinct semantic phenomena, and in particular an analysis of the geometric features of subspaces which correlate with different measures of the conceptual interrelations between lexical representations. This meta-analysis will serve to test my hypothesis that different statistical features of an appropriately contextualised semantic space map to different semantic

[1] Copies of the instructions, along with the data itself, can be found at `www.cs.technion.ac.il/ gabr/resources/data/wordsim353/wordsim353.zip`.

[2] Instruction and data are at `https://staff.fnwi.uva.nl/e.bruni/MEN`.

phenomena, and the corresponding claim that context sensitive representations can capture various semantic features as dynamic properties in a single subspace. Finally, the analysis of the different geometric correlates of relatedness and similarity lends itself to a consideration of the way in which the frames within which humans evaluate semantic relationships may themselves be contextual.

3.1 An Experiment on Relatedness

Standard distributional semantic models have generally tended to capture semantic relatedness over similarity in terms of the proximity between semantic representations. This point, evidenced by the stronger results achieved on relatedness tests by statistical models, can be seen clearly by imagining the contexts in which words such as *good* and *evil* or *day* and *night* might be expected to regularly occur: there is no serious case to be made that the meaning of a sentence would not be significantly changed by toggling these word pairs in actual sentences (they are closer to being antonyms than to being synonyms), but it is equally reasonable to guess that these words will generally have similar co-occurrence profiles. Examples of corpus derived, distributional semantic type models that have performed well on on word relatedness evaluations include the work of **?** and **?**, both of whom have applied vector building techniques that exploit Wikipedia page labels to enhance the conceptual knowledge inherent in their lexical representations. **?** similarly enhance neural word embeddings derived from co-occurrence observations with synonymy information extracted from WordNet. And **?** use recursive neural networks to actually move to a level of linguistic abstraction below the word itself, modelling the morphology and the corresponding composition of words based on morphemes as a productive element in predicting relatedness between words. The overall import of this literature is that there is scope for using corpus analytic techniques to build lexical representations that do a good job of capturing semantic relatedness.

Nonetheless, there may be some advantages to identifying context specific subspaces based on an analysis of word pair inputs. For instance in cases where one of the words being compared has multiple senses, the selection of mutually relevant co-occurrence dimensions under the JOINT and ZIPPED techniques might offer a degree of disambiguation. Beyond this, I hypothesise that similar measures to the ones that have proved productive for static vector space models, so, in particular, measures of cosine similarity between word-vectors, anchored at the origin as well as at the generic points of the space, should be indicative of semantic relatedness. I further predict, following on the results reported earlier in this chapter on the relationship between the norm of vectors in contextualised subspaces and conceptual entailment, that measures involving the distance of word-vectors from the origin will also correlate positively with relatedness, and here my subspaces, with their sense of interior and exterior, centre and periphery, should have an advantage.

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.666	0.681	0.698	0.728	0.704	0.698	0.700	0.709
INDY	0.671	0.676	0.702	0.707	0.703	0.712	0.715	0.729
ZIPPED	0.642	0.674	0.699	0.698	0.652	0.678	0.716	0.717
SVD	0.521	0.618	0.690	0.728	0.527	0.663	0.722	0.742
SG	0.549	0.639	0.696	0.701	0.544	0.635	0.705	0.710
CBOW	0.557	0.648	0.700	0.695	0.584	0.663	0.716	0.716

Table 3-A: Spearman's correlations for word ratings output by a linear regression model of the WordSim data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

3.1.1 Relatedness: Methodology and Model

In order to test the ability of my statistical methodologies to likewise model relatedness, I build JOINT, INDY, and ZIPPED subspaces using each of the 353 word pairs in the WordSim data as input. I project subspaces of 20, 50, 200, and 400 dimensions, extrapolated from base spaces built using 2x2 and 5x5 word co-occurrence windows. For each subspace, I extract the geometric features listed in the previous chapter in Table **??** and use these as the independent variables of a linear regression, taking the WordSim rating of the word pair used to generate the subspace as the dependent variable. The relatedness ordering of word pairs inherent in the scores assigned by the regression are then compared to human WordSim ratings in terms of Spearman's correlations, as is standard practice in the NLP literature. Results from my model are compared with results from singular value decompositions of my base space using comparable parameters, as well as `word2vec` skip-gram and bag-of-words models, again using commensurable parameters.

Results are reported in Table 3-A (and all correlations are statistically significant with $p < .001$). The first thing to note is that the best performance overall is achieved by the 5x5 word window, 400 dimensional version of the SVD factorisation of my base space (though the difference between this correlation and the slightly lower correlation achieved with the same parameters for the INDY dimension selection technique is not significant, with $p = .356$ based on a Fisher transformation). More generally, the 5x5 word co-occurrence window versions of all models tend to perform more strongly on this task than the 2x2 versions, suggesting that semantic relatedness is a property of the broader sentential context in which a word occurs rather than just the immediate syntagmatic tendencies of a word.[3] It is also notable that my context sensitive methods outperform the static models at lower dimensionality (and here the difference is significant, with $p < .005$ in a comparison between the JOINT 5x5 window, 20 dimensional correlation and the corresponding result for the CBOW model). It seems that the contextually selected dimensions are initially all more informative about relatedness than the degree of general variance captured in lower numbers of dimensions using either factorisation or neural

[3] **?** discusses **?**'s (**?**) semiotic notions of *syntagm* (the way that words are composed into meaningful utterances) and *paradigm* (the way that words are comparable and potentially interchangeable units of meaning) in the context of distributional semantics.

modelling techniques.

In terms of comparing between my dimensional selection techniques, the JOINT and INDY techniques perform somewhat comparably, with the INDY technique doing a bit better in the 5x5 spaces. The strong performance of the JOINT technique in the 2x2 models at 400 dimensions seems anomalous. The joint technique in particular should begin to run out of useful dimensions to select between words as the dimensionality of the spaces scale up, and so would be expected to follow a similar trend as with the ZIPPED subspaces, where results begin to tail off after 200 dimensions—and this effect should be more prominent in the 2x2 models, where there is less overall co-occurrence information available. It's likewise interesting that the ZIPPED technique offers consistently lower correlations, particularly considering that this technique was conceived as something of a hybrid between the comprehensive JOINT approach and the independent INDY approach. It would seem, then, that the dimensions most predictive of semantic relatedness are either those which are substantially informative about both words being compared, or those which are highly informative about one word and only incidentally informative about the other, to the exclusion of the middle ground of dimensions that are highly informative about one word and at least marginally informative about another. The conclusion to draw here is that the JOINT and INDY spaces are identifying relatedness in two different capacities: in the case of the former, the degree of proximity between two points with fairly high values is being captured, while in the case of the latter the extent to which there is some degree of overlap (or, alternatively, the extent of the orthogonality) between the salient co-occurrence features is being exploited.

3.1.2 Relatedness: Comparing to Other Models

It must at this point be noted that the models described above are instances of fitting the output produced by my methodologies, and so they should not be construed in some sense as solutions

In order to get a sense of what's actually happening in these models, I next produce Spearman's correlations between the WordSim data and each of the features of different subspaces independently. The top five features for 400 dimension JOINT, INDY, and ZIPPED spaces generated using 2x2 word co-occurrence windows are reported in Table 3-B. Here a strong correlation between the most predictive of the JOINT and ZIPPED subspaces is evident, and this makes sense: as these types of subspaces increase in dimensionality, the possible combinations of co-occurrence dimensions with non-zero values for both input word-vectors decreases, so the subspaces themselves begin to converge. The features selected here tend to involve the mean norm of the input word-vectors, so the prominence of these vectors in the spaces that are jointly informative about both of them is clearly positively correlated with relatedness between the two terms.[4] In other words, related words tend to share strong PMI values with a number of co-occurrence dimensions—hardly a surprising finding, and in line with the results indicating the powerfulness of norm measures revealed in the proof of concept outlined earlier in this chapter.

[4] There is clearly a high degree of multi-colinearity at play between these top independent features, and this will be addressed below.

JOINT		INDY		ZIPPED	
$\angle AMB$	0.645	$\angle ACB$	0.721	$\angle AMB$	0.636
$\angle ACB$	0.636	$\angle AMB$	0.703	$\angle ACB$	0.607
$\mu(A,B)/M$	0.604	$\angle A'C'B'$	0.663	$\mu(A,B)$	0.603
$\mu(A,B)$	0.604	$\angle A'X'B'$	0.634	$\angle A'M'B'$	0.593
$\mu(A,B)/C$	0.603	$\angle AOB$	0.634	$\angle A'X'B'$	0.587

Table 3-B: Independent Spearman's correlations with WordSim data for top five features of each subspace type for 5x5 word co-occurrence window, 400 dimension subspaces.

Much more interestingly, though, an altogether different set of top features emerges for the INDY subspaces. Here, angular measures are more predictive of relatedness across the board, with the measure $\angle ACB$, the angle of the word-vector points A and B at the vertex of the central point C, being independently more predictive than many of the combined features in lower dimensional spaces. It should be noted at this point that angles are measured in terms of cosine, so a strong positive correlation indicates that angles become smaller in terms of degrees as words become more related. In fact, the measure $\angle AOB$ is just the cosine similarity of the word-vectors, so here the INDY subspaces are seen aligning somewhat with the standard approach from static spaces. The strong correlations between small angles with the generic points of the space ($\angle ACB$ and $\angle AMB$) as well as the normalised version of these points ($\angle A'C'B'$ and $\angle A'X'B'$) emphasises the point that related words tend to select subspaces where their word-vectors are relatively close to each other compared to their proximity to the maximal, central, and mean vectors in their INDY subspace.

So, where the JOINT and INDY subspaces provide a basis for correlation between norms and relatedness, the INDY subspaces evidently create a similar axis of correlation between angles and relatedness. In order to delve deeper into the models learned from the geometric features of these spaces, I next discover an optimally predictive and uncorrelated combinations of five features for modelling relatedness in each type of subspace. Treating this process as a breadth-first search of possible linear combinations of features to be fed to a regression model, I begin with each independent feature and then concatenate additional features with the constraint that each added feature must not have a *variance inflation factor* **?** of greater than 10 with an existing chain of features. So, if R_i^2 is the coefficient of determination of adding independent variable i to an $i-1$ linear model, then the addition is only considered if it satisfies the inequality $1/(1-R_i^2) < 10$. This constraint serves two purposes. First, it eliminates multicolinearity in the combinations of features learned by the model; this, in turn, results in a combination of features which is optimally informative about the information contained in the geometry of a type of subspace and also in model coefficients which are interpretable in terms of their scale. Second, it makes a potentially very large state space of feature vectors computationally tractable by eliminating a good proportion of possible combinations of features at each level of the search tree. So, for instance, in the case of the 2x2 word, 400 dimensional JOINT subspace, the state space of 33,390,720 five feature long combinations selected from 34 different features becomes a space of just 194,481 combinations.

The top five features for each dimensional selection technique, applied to the 2x2 word

JOINT ($\rho = 0.668$)		INDY ($\rho = 0.728$)		ZIPPED ($\rho = 0.668$)	
$\angle AMB$	1.463	$\angle ACB$	1.833	$\angle AMB$	1.363
$\overline{A'M'} : \overline{B'M'}$	0.795	$\angle A'C'B'$	-0.554	$\mu(A, B)$	0.743
$\overline{A'X'} : \overline{B'X'}$	-0.675	$\overline{A'M'} : \overline{B'M'}$	-0.123	$\overline{A'X'} : \overline{B'X'}$	0.491
$\overline{AB}$	0.274	$\mu(AB)/X$	-0.101	$\mu(AB)/C$	-0.424
$\overline{AM} : \overline{BM}$	-0.230	$A : B$	-0.066	$\overline{AC} : \overline{BC}$	-0.154

Table 3-C: The optimal combination of five non-correlated features for a linear regression modelling WordSim data for 5x5 word co-occurrence window, 400 dimensional subspaces projected using each dimensional selection technique.

co-occurence window base space and returning 400 dimensional subspaces, are listed in Table 3-C, with the Spearman's correlation achieved by each combination of features listed in parentheses next to the technique labels. The first thing to note here is the variety of features evident throughout this table: angles, distances, means of distances, and ratios of distances are all to be found, involving measurements in both the extents of space and between the normalised versions of the two word-vectors and the three generic vectors. Next it is interesting to see that, once again, different features prove most predictive in different types of subspaces. In particular, the mean norm values, represented as $\mu(A, B)$, continue to correlate positively with relatedness in both JOINT and ZIPPED subspaces, suggesting that with sets of collectively informative dimensions, consistently strong values for both word-vectors indicate a high degree of relatedness. In the case of the INDY subspaces, on the other hand, the angle $\angle ACB$ continues to be highly predictive of relatedness, with smaller angles at the vertex of the central vector indicating a higher degree of relatedness.

There are, however, also some interesting new consistencies which emerge between subspace types. For both the JOINT and INDY subspaces, for instance, the ratios of distances between each word-vector and some of the generic points are predictive of relatedness. For the INDY subspaces, the correlation with ratios of distances from the central point C and the normalised version of this point C' is negative; since the ratio measure always divides the smaller value by the larger, this means that more lopsided proximities between word-vectors and the line going through the centre of a subspace tend to actually correlate with relatedness in subspaces where each dimension is selected for its salient co-occurrences with just one of the words being analysed. In the case of the JOINT subspaces, the negative correlation with $\overline{A'C'} : \overline{B'C'}$ is offset by a positive correlation with $\overline{A'M'} : \overline{B'M'}$, the ratio of the distances from each normalised word-vector to the normalised mean vector. So here it turns out that, when words are more closely related, the typical distances between each word-vector and the central line tend to be more lopsided, but at the the typical distance between each word-vector and the vector of mean values across a subspace, which in a certain regard delineates the true statistical centre of a subspace, tend to be more similar.

This last observation serves as a reminder that these subspaces are not necessarily composed of dimensions with uniform statistical properties. On the contrary, referring back to the analysis of mean and maximum values in Table **??**, we recall that there tends to be a good deal of variance in both of these statistics, and so we can presume that subspaces will exhibit some degree of distortion. This is reflected in the salience of features in both the JOINT and INDY subspaces which don't involve the word-vectors themselves. In particular the negative correlation between $\angle COX$ and relatedness in JOINT subspaces means that wider angles between a vector with uniform values and a vector of maximum dimensional values indicate relatedness between the words that select those dimensions, so related words tend to jointly select dimensions with greater variance in their maximum values. Maximum values again play a role in predicting relatedness in INDY subspaces, where simply the norm of the vector of maximum values X correlates positively with relatedness. Since higher PMI values will tend to occur along dimensions where the frequency of the corresponding co-occurrence term is lower, we can infer that words with a tendency to be related tend to have high PMI values with less frequent co-occurrence terms.

This last observation might at first seem counter-intuitive: can it really be the case that some sets of dimensions just tend to be more characteristic of related words? And can simply the frequency with which some word occurs to some extent predict the likelihood of a person thinking that word is related to other words? I claim that the answer to these questions is "yes". There are relatively simple statistical properties that correlate in logical ways to some of the cognitive

This claim will be explored further below in Section 3.2.1, and then in more detail in the following chapter exploring my methodology's capacity for classifying figurative language. First, though, I will present results on a similar experiment involving similarity rather than relatedness.

3.2 An Experiment on Similarity

Where relatedness has been a fruitful target for statistical semantic modelling, word similarity has typically been the domain of models endowed with a degree of encyclopedic knowledge about the world. A Spearman's correlation of $\rho = 0.76$ with the human evaluations of the SimLex data, a result comparable with inter-annotator agreement, is achieved by **?** using a statistical model enhanced with a weighted graph of conceptual relationships extracted from the 4lang *conceptual dictionary* (**?**). **?** similarly use a combination of statistical and knowledge based models, treating the outputs of individual models developed by various researchers as the independent variables of a range of regression models, achieving correlation of $\rho = 0.658$ in the case of the best performing model. Statistical approaches, on the other hand, have included models such as the one described by Schwartz et al. (2015), which combines `word2vec` word-vectors with vectors of syntagmatic *systematic patterns* of co-occurrence which the authors predict will be particularly indicative of semantic similarity, producing a correlation of $\rho = 0.563$. And the first shot at the SimLex data, presented in the same paper that presented the dataset itself (**?**), achieves

window	2x2				5x5			
dimensions	20	50	200	400	20	50	200	400
JOINT	0.414	0.444	0.471	0.459	0.404	0.412	0.425	0.429
INDY	0.411	0.445	0.481	0.503	0.391	0.429	0.462	0.490
ZIPPED	0.425	0.446	0.480	0.471	0.400	0.406	0.430	0.446
SVD	0.235	0.274	0.375	0.423	0.218	0.255	0.353	0.380
SG	0.232	0.273	0.337	0.379	0.215	0.252	0.322	0.355
CBOW	0.245	0.290	0.367	0.404	0.247	0.290	0.372	0.406

Table 3-D: Spearman's correlations for word ratings output by a linear regression model of the SimLex data for various subspace types and model parameters, compared to the correlations for cosine similarities output by static models using comparable parameters.

JOINT		INDY		ZIPPED	
$\mu(A,B)/C$	0.377	$\angle ACB$	0.398	$\mu(A,B)/M$	0.361
$\mu(A,B)/M$	0.376	$\angle AMB$	0.375	$\mu(A,B)/C$	0.361
$\mu(A,B)/X$	0.356	$\angle A'X'B'$	0.357	$\mu(A,B)/X$	0.343
$\angle AMB$	0.349	$\angle A'C'B'$	0.351	$\angle AMB$	0.342
$\angle ACB$	0.349	$\angle AOB$	0.333	$\angle ACB$	0.325

Table 3-E: Independent Spearman's correlations with SimLex data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces.

3.2.1 Comparing the Two Phenomena

[5]

3.2.2 Generalising the Models

3.2.3 Frames of Similarity

?, in his psychologically motivated reflections on the geometry of similarity, observes that relationships of similarity are fundamentally not symmetric:

(for comparison, see where the pair (*dentist, colonel*) falls in the space of Figure 3.1, where it is one of the very few pairs considered to be marginally more similar than related)

[5] Intriguingly, when identical words are given as input, they are rated as being very related and very dissimilar. The latter outcome is obviously an imperfection, but it also reveals the extent to which the models of each type of semantic phenomenon are making use of different geometric features, or the same features in opposite ways.

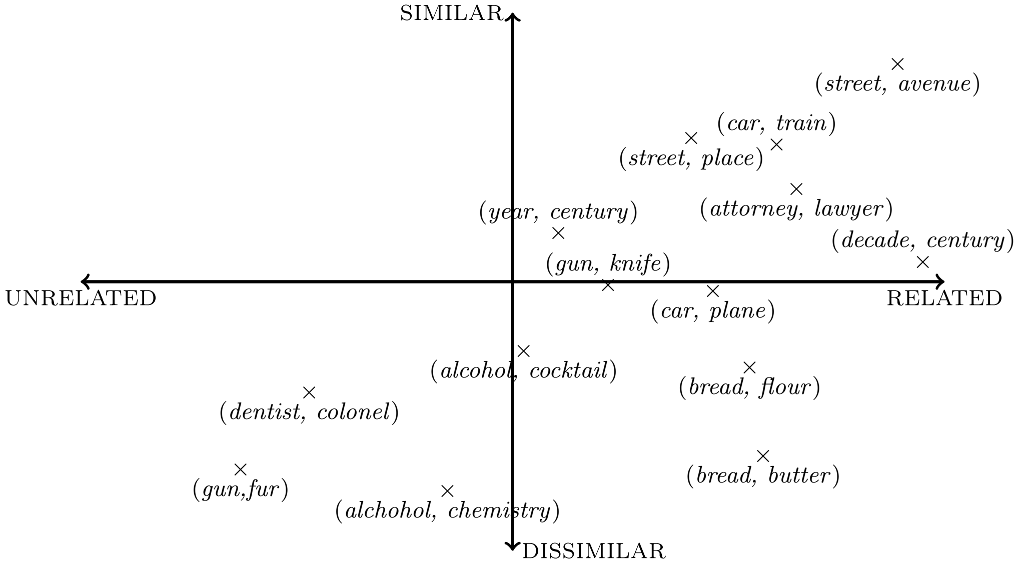
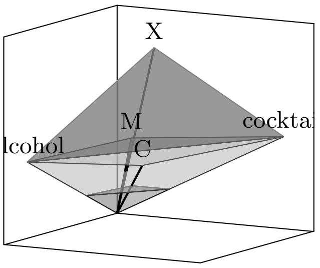
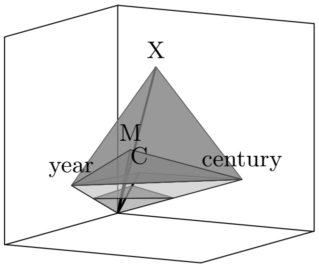


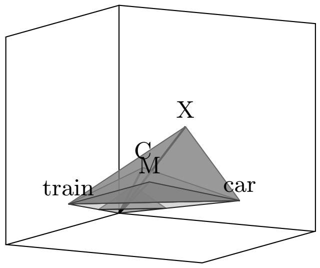
Figure 3.1: Noun pair scores along axes of relatedness and similarity as returned by a model built from features of 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.



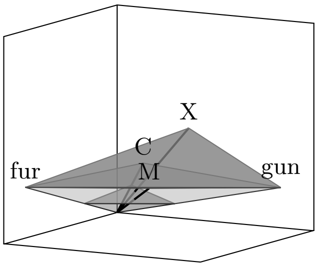
relatedness = 5.129
similarity = 3.714



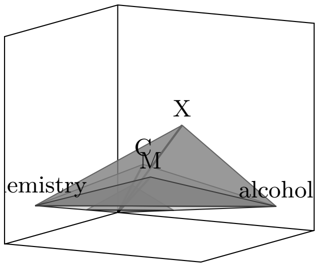
relatedness = 5.530
similarity = 5.901



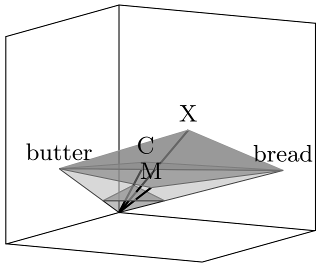
relatedness = 8.057
similarity = 7.550



relatedness = 1.852
similarity = 1.512



relatedness = 4.247
similarity = 1.109



relatedness = 7.901
similarity = 1.760

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Agres, K., McGregor, S., Purver, M., and Wiggins, G. (2015). Conceptualising creativity: From distributional semantics to conceptual spaces. In *Proceedings of the 6th International Conference on Computational Creativity*, Park City, UT.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014a). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9:241–346.

Baroni, M., Dinu, G., and Kruszewski, G. (2014b). Don't count, predict! In *ACL 2014*.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* MIT Press, Cambridge, MA.

Clark, S. (2015). Vector space models of lexical meaning. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*, pages 493–522. Wiley-Blackwell.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167.

Dummett, M. (1981). *Frege: Philosophy of Language.* Duckworth, London, 2nd edition.

Erk, K. and Smith, N. A., editors (2016). *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Berlin, Germany.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Bradford Books.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611.

Gärdenfors, P. (2000). *Conceptual Space: The Geometry of Thought.* The MIT Press, Cambridge, MA.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Houghton Miffline, Boston.

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.*

Hill, F. and Korhonen, A. (2014). Learning abstract concept embeddings from multimodal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 255–265.

Kiela, D., Hill, F., and Clark, S. (2015). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2048.

Kottur, S., Vedantam, R., Moura, J. M. F., and Parikh, D. (2016). Visualword2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4985–4994.

Lapesa, G. and Evert, S. (2014). A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc.

McGregor, S., Agres, K., Purver, M., and Wiggins, G. (2015). From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence.*

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

Mikolov, T., tau Yih, W., and Zweig, G. (2013c). 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*, pages 246–251.

Milajevs, D., Sadrzadeh, M., and Purver, M. (2016). Robust co-occurrence quantification for lexical distributional semantics. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 58–64, Berlin, Germany. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing.*

Polajnar, T. and Clark, S. (2014). Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 230–238.

Schwartz, R., Reichart, R., and Rappoport, A. (2015). Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 258–267.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

van der Velde, F., Wolf, R. A., Schmettow, M., and Nazareth, D. S. (2015). A semantic map for evaluating creativity. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, pages 94–101.

von Uexküll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds. In Schiller, C. H., editor, *Instinctive Behavior: The Development of a Modern Concept*, pages 5–80. International Universities Press, Inc., New York City, NY.

Widdows, D. (2004). *Geometry and Meaning*. CSLI Publications, Stanford, CA.