[[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

Conceptual Clusterings, Similarity, and Relatedness

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 much. In this chapter, I will now 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 semantics based on statistics about word co-occurrences. They encompass two different experimental set-ups and corresponding varieties of data, one of which has been designed specifically for the purpose of testing my ideas and one of which involves an assortment of data used pervasively by computational linguistics interested in semantic models.

The first experiment, presented as a proof of concept, involves using multi-word phrases as input and evaluating the methodology's capacity for building subspaces where words associated with the conceptual category denoted by the input term can be reliably discovered. This experiment expands upon the notion of proto-conceptual spaces outlined in the previous chapter, considering whether the word vectors that populate regions of subspaces are characterised by a certain categorical coherence. In the case of the data explored here, the experiment is specifically set up to feel out the contextual capacity of my methodology and compare it to a standard generic semantic space. The question asked is whether the shifts from subspace to subspace based on particular input yield productive alterations in the way that words both cluster and emerge from the melange of word-vectors that circulate around my base model.

The second experiment moves into more familiar computational linguistic territory, using some well-travelled datasets to examine the methodology's capacity for identifying two related but distinct semantic 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.

3.1 A Proof of Concept

In this section, I present the first experiment performed using my contextually dynamic distributional semantic model. The gist of this experiment is to take a word pair representing a compound noun – for instance, *body part* – and see if my methodology can use the word pair to contextually generate a space where other words conceptually related to that compound noun can be found in a systematic way. This is conceived of as an entailment task, in that I will attempt to find phrases considered to be categorical constituents of the concept represented by the word pair, taking the WordNet lexical taxonomy as a ground truth. There is a scholastic back story here.

An early version of this experiment was reported in Agres et al. (2015). That first effort arose out of a question posed by a colleague regarding the feasibility of using a statical NLP technique for generating categorical labels that could be used to evaluate computational creativity in a domain specific way (for a psychological perspective on the difficulty of generating such terms in an objective way using human subjects, see van der Velde et al., 2015). So, for instance, given a creative domain such as MUSICAL CREATIVITY, could a distributional semantic model generate terms that are reliably relevant to the concept denoted by that phrase, rather than the potentially disparate properties independently associated with MUSIC and CREATIVITY? Intuitively there seems to be little reason to hope that the space halfway between these points in a general semantic space would somehow adequately represent the properties of the overall concept. The early work explored the dimensions contextually selected by analysing the co-occurrence features of word-vectors corresponding to inputs along the lines of the expository results presented anecdotally in Chapter **??**, but without any rigorous evaluation.

Reviewer responses to a subsequent journal article (McGregor et al., 2015), designed as a more thorough introduction of the methodology, inspired a computationally oriented mode of evaluation. The experiment that has emerged involves attempting to recapitulate taxonomical conceptual relationships from the WordNet database (Fellbaum, 1998). Wordnet is a lexical taxonomy of *synsets*, basically semantic word senses, arranged into a hierarchy of entailment relationships, with each synset associate with a number of *lemmas*, word types indexed by that synset according to human annotators. This experiment takes as input instances of synsets labelled by compound noun phrases and seeks to output as many of the lemmas listed associated with synsets that are hyponyms of the input synset. So, for instance, the synset body part has a hyponym EXTERNAL BODY PART, which has a hyponym EXTREMITY, which has a synset LIMB, which has a synset LEG associated with the lemma *leg*, and so *leg* would be considered a positive output for the input *body part*.[1]

3.1.1 Experimental Set-Up

12 of the top synset labels consisting of compound noun phrases are extracted from WordNet. These labels are extracted through a breadth first traversal of the tree of noun synsets, selecting the highest 12 synsets with multi-word labels with the constraint that none of the 12 can be parent nodes of any of the others: in this way, 12 distinct, non-overlapping conceptual categories are choosen. The experimental vocabulary is considered to be the intersection of the list of all WordNet noun lemmas associated with the vocabulary of my model (the 200,000 most frequent word types in Wikipedia), resulting in a total vocabulary of 32,155 words. The lemmas associated with all the hyponyms of each synset are extracted and grouped, and these words become the target words for my models' output. The 12 synset labels are itemised in Table 3-B.

With the target output established, the terms labelling a given synset are passed to my model as contextual input, with the corresponding word-vectors serving as the basis for dimensional selection using the JOINT, INDY, and ZIPPED techniques as outlined in Chapter **??**. Here, the base space generated using a 5x5 word co-occurrence window is used, and 200 dimensional subspaces are returned; variations of these parameters will be tested in subsequent experiments. The subspaces returned by each of these techniques are explored to return the top terms using both of the procedures outlined in Chapter **??**: the terms closes to the mean point between the input word-vectors in a subspace are returned, and the terms furthest from the origin – the terms with the largest norm – in a given subspace are returned. The top 50 terms found in a subspace each according to each measure are returned, as well as the top terms up to a limit n where n is the total number of lemmas associated with the target multi-word label. Accuracy scores for each of these sets of output are computed, so the total number of positive matches for hyponyms of the input synset out of the top 50 and top n terms returned.

As a point of comparison, results are likewise returned from two different `word2vec` models, one using the skip-gram methodology and one using the bag-of-words methodology, as described in Chapter **??**. In line with the subspaces generated using my methodology, 200 dimensional models are used, and these models are built across 10 iterations of the corpus, using a 5x5 word co-occurrence window, applying a negative sampling rate of 10 and an initial learning rate of 0.025, as discussed in Chapter **??**. Here the top terms in terms of proximity by cosine similarity to the mean point between the word-vectors associated with the input terms are returned, again taking the top 50 and top n for each input.

[1] In keeping with the convention used elsewhere in this thesis, synset labels will be presented in small caps and lemmas will be presented in italics.

		JOINT		INDY		ZIPPED			
		norm	dist	norm	dist	norm	dist	SG	BoW
top-50	accuracy	0.292	0.208	0.240	0.189	0.273	0.199	0.247	0.270
	ratio	10.304	6.129	7.731	5.270	8.625	5.719	6.733	7.168
full	accuracy	0.235	0.160	0.198	0.149	0.210	0.153	0.081	0.079
	ratio	4.967	3.525	3.967	2.997	4.290	3.221	2.397	2.551

Table 3-A: Average accuracy scores and average ratio of accuracy to baseline for reconstructing the lemmas entailed by 12 different multi-word WordNet synsets, for both the top 50 terms returned by models and the full set of terms returned up to the number of lemmas associated with each input.

3.1.2 Results and Analysis

Results for the set-up described in the previous section can be found in Table 3-A, with both the average accuracy scores and the average ratio of model accuracy to baseline reported. Results for both the norm and distance from mean point methods are reported for subspaces derived using the JOINT, INDY, and ZIPPED dimension selection techniques, followed by results for the skip-gram and bag-of-words `word2vec` techniques. The first thing to note about these results is that all of the results are substantially above the baseline: the average ratios of model accuracy to the baseline (the likely accuracy achieved by randomly choosing words from the vocabulary for each input) are all above 2.5, and are above 3.2 for all of my methodologies. So it is clear that all these techniques are generating semantically significant relationships between word-vectors.

Results across the board are strongest for the JOINT dimension selection technique applying the norm measure for returning output: in these subspaces selected by choosing dimensions with high PMI values across all contextual inputs, word-vectors that are far from the orgins – and that therefore likewise tend to have high values across all these dimensions – are most characteristic of the conceptual category indicated by the input. This is not surprising. Results for the norm measure applied to ZIPPED and INDY type subspaces follow in kind, with intermediary performance from the in-between ZIPPED technique, where all dimensions bear at least some tendency for co-occurrence with the input terms, and then another step down for the INDY subspaces. In all cases the norm measure outperforms the two `word2vec` results.

More surprising is the distinction between the strong performance of the norm measures and the less impressive performance of the mean point measure. In the case of accuracy among the top 50 terms returned by each model, my methodologies results using this Euclidean measure consistently fall short of the `word2vec` techniques. It would seem, then, that in the subspaces returned by my models, proximity to the input word-vectors is not in itself an indicator of categorical inclusion in the conceptual space traced by the intersection of the correspond contextual input terms. Upon further consideration, there is a plausible explanation for this: revisiting the outputs for subspaces projected using denotations of animals as input, reported last chapter in Tables **??** and **??**, the norm measure produced specialised terms such as *chital* and *poodle*, while the distance

		top-50			full		
	baseline	norm	dist	BoW	norm	dist	BoW
psychological feature	2.39	0.240	0.660	0.400	0.401	0.417	0.102
causal agency	0.177	0.000	0.140	0.180	0.125	0.170	0.043
human action	0.156	0.180	0.460	0.480	0.300	0.346	0.116
animate being	0.044	0.020	0.060	0.020	0.030	0.031	0.006
cognitive content	0.043	0.360	0.260	0.300	0.168	0.188	0.050
mental object	0.043	0.120	0.240	0.180	0.130	0.188	0.053
physical process	0.035	0.520	0.260	0.200	0.205	0.138	0.065
social group	0.031	0.080	0.220	0.380	0.075	0.114	0.064
body part	0.025	0.760	0.120	0.220	0.407	0.080	0.087
taxonomic category	0.024	0.460	0.180	0.540	0.147	0.026	0.164
physiological condition	0.020	0.640	0.160	0.280	0.365	0.099	0.139
woody plant	0.012	0.120	0.060	0.060	0.143	0.127	0.062

Table 3-B: Item-by-item accuracy results for the entailment experiment run on WordNet synsets, reported for the norm and distance metrics using the JOINT technique as well as `word2vec's` bag-of-words method.

measure generated relevant but not always categorical terms such as *wild*, *giant*, and *golden*. To give an example from the data used for this experiment, top-50 results from the JOINT distance measure returned for the input (*body, part*) include words like *portion*, *upper*, *shape*, and *whole*, while the results from PHYSICAL PROCESS include *method*, *complex*, and *affect*—so, terms that are conceptually relevant to the target domain but are not strictly part of the category BODY PART. We might characterise this trend in terms of a distinction between words which denote semantic *relatedness* versus *similarity*, a topic which will be addressed in depth in the next section.

Focusing on the accuracy of the results returned by the models up to the full length of each target set of lemmas, here results are weaker all around, which is not particularly surprising: as we move away from the regions where we expected to see the highest degree of conceptual consistency, mismatched terms begin to creep into the results. It is notable, though, that my methodologies outperform the neural network based models across the board, especially for the norm based measures but also in the case of this larger sample of the respective semantic spaces for the distance based measures. In fact, the stronger relative performance for the distance measure in these expanded regions of each type of subspace makes sense, since, as the norms measure moves closer to the origin in search of output and the distance measure likewise expands from the locus of its mean point, the results output by each measure will increasingly overlap (an overlaying of Figures **??** and **??** will illustrate this phenomenon). But the main point to take here is that, in the case of my methodologies, there is clearly a more persistent conceptual organisation to the space. As we expand from any point in the static type of semantic model generated by `word2vec`, we will undoubtedly begin to encounter the vagary and the messiness inherent in language and problematic for fixed lexical relationships. My methodologies, on the other hand, afford the *ad hoc* construction of semantic spaces which afford the situational corralling of the looseness and ambiguity inherent in a dynamic lexicon.

Table 3-B presents accuracy rsults for each of the 12 conceptual categories targeted by this experiment, focusing on the two measures applied to JOINT type subspaces as well as the bag-of-words version of the `word2vec` methodology. It's particularly pleasing to see my methodology handling the ambiguity inherent in the inputs (*body, part*) and (*physical, process*) so well as it finds the relevant terms very far from the origin, while, as discussed above, the distance measure falls short here, presumably because it is finding terms that are related to the input rather than terms that are entailed by it. On the other hand, the distance measure does quite well for inputs such as (*psychological, feature*) and (*human, action*). A pitfall for the norm measure and the bag-of-words method is that they both seem to have identified a region of PSYCHOLOGICAL [THRILLER] FEATURE [FILM], yielding outputs such as *slasher*, *offbeat*, and *blockbuster*, so there is clearly still scope for ambiguity here even with a degree of context. It's interesting to observe how the norm measure manages to recover from this category error as it returns more results, whereas the bag-of-words method evidently wanders further off topic. That said, the bag-of-words results are impressive, at least in the top 50 outputs, for the inputs (*social, group*) and (*taxonomic, categories*), arguably instances where the context is already somewhat evident with one of the two inputs.

These are, on the whole, promising results for my methodology. They illustrate its ability to delineate a context specific subspace based on a conceptually targeted input and then discover regions within this space that evidence a degree of conceptual inclusion. Furthermore, the regions discovered seem to be relatively well defined, with a lesser degree of dithering away from the top or centre of the regions compared to a standard static semantic model. On the other hand, the outputs from these regions are marked by an different kind of ambiguity than polysemous word senses: there is a confusion between words which denote entities entailed by the input, and words which simply relate to the input. The next section will expose the methodology to a group of datasets that have already been broadly reported in the computational linguistic literature, with the objective of establishing precisely the ability of context sensitive models to make distinctions between similarity and relatedness.

3.2 Relatedness and Similarity

A fundamental objective for a general semantic model is a mechanism for representing 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" (Rubenstein and Goodenough, 1965, p. 628) in each pair, or "how strongly these words are related in meaning," (Yang and Powers, 2006, p. 124). Finkelstein et al. (2002) 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";[2] Bruni et al. (2012) used only the term *relatedness* in their instructions, with no mention of *similarity*.[3] Faruqui et al. (2016) 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. Agirre et al. (2009) 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 Hill et al. (2015) 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 phenomena, and the corresponding claim that context sensitive representations can cap-

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

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

ture 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.

3.2.1 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, 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.

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

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

Table 3-C: 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.

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-C. The first thing to note is the very strong performance of the SVD version of my base space

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-D. 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.

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.

[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	
$\mu(A, B)$	0.609	$\angle ACB$	0.683	$\mu(A, B)$	0.611
$\mu(A, B)/C$	0.604	$\angle AMB$	0.654	$\mu(A, B)/X$	0.603
$\mu(A, B)/X$	0.603	$\angle A'C'B'$	0.600	$\mu(A, B)/C$	0.598
$\mu(A, B)/M$	0.602	$\angle AOB$	0.594	$\mu(A, B)/M$	0.596
$\angle ACB$	0.574	$\angle A'X'B'$	0.571	$\angle AMB$	0.566

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

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

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 WordSim data for top five features of each subspace type for 2x2 word co-occurrence window, 400 dimension subspaces.

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.

3.2.2 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.

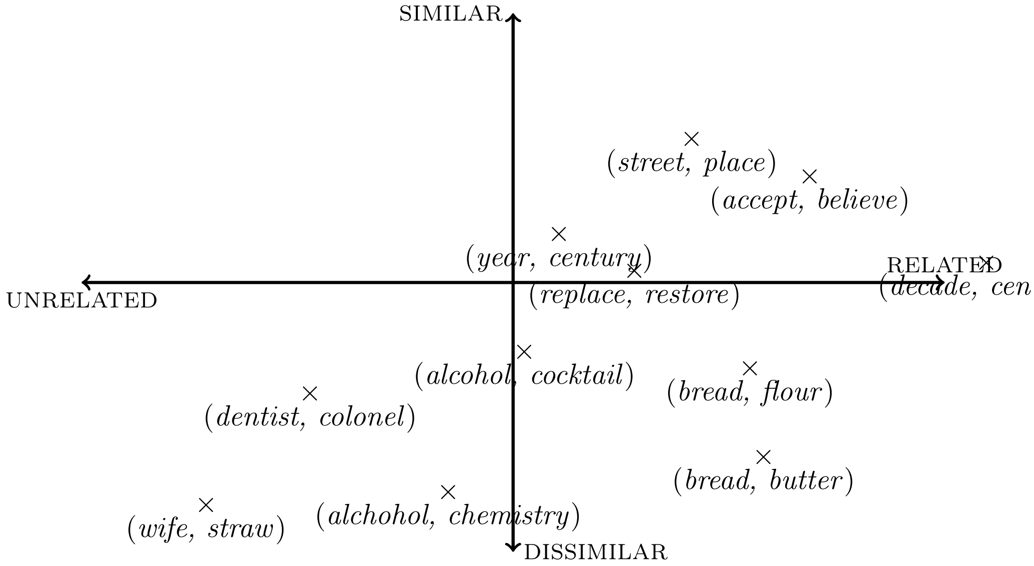

Figure 3.1: Word pair scores along axes of relatedness and similarity as found in 2x2 word co-occurrence window, 400 dimensional, INDY type subspaces.

3.2.3 Comparing the Two Phenomena

3.2.4 Generalising the Models

3.2.5 Frames of Similarity

Tversky (1977), 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)

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27.

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.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 136–145.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

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

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transanction on Information Systems*, 20(1):116–131.

Hill, F., Reichart, R., and Korhonen, A. (2015). Simlex-999: Evaluating semantic models with genuine similarity estimation. *Computational Linguistics*, 41(4):665–695.

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*.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communincations of the ACM*, 8(10):627–633.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

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.

Yang, D. and Powers, D. M. W. (2006). Verb similarity on the taxonomy of wordnet. In *3rd International WordNet Conference*, pages 121–128.