Stage Two Report:
A Statistical Language Model for the Generation of Figurative Language

by

Stephen McGregor

An outline of a thesis to be submitted to the University of London
for the degree of Doctor of Philosophy

Department of Electronic Engineering
Queen Mary, University of London
United Kingdom

September 2015

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 computational 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.

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 being situated 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, often complemented by additional 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 1

A Computational Implementation of Context Sensitive Distributional Semantics

In the previous chapter, I laid down the theoretical groundwork for a distributional semantic methodology for dynamically establishing perspectives on statistical data about language use. In this chapter, I'll describe the technical details for building a computational implementation of such a methodology. The objective of this implementation is to establish a rigorous procedure for generating subspaces of word vectors, based on observations of word co-occurrences in an underlying corpus, the geometries of which are semantically productive in particular contexts. This will involve three steps:

1. The selection, processing, and analysis of a large scale textual corpus in order to create a high dimensional base space of co-occurrence statistics;

2. The development of techniques for selecting lower dimensional subspaces based on some sort of contextualising input;

3. The exploration of the geometry of the projected subspaces in search of semantic

correlates.

The following three sections will pursue each of these aspects of a technical implementation in turn. The end result is effectively a mapping from text as raw data to geometry as semiotic generator. A fourth section will describe an alternative, general interpretation of the statistical data which underwrites my models and additionally offer a brief overview of another distributional semantic methodology, all to be used as a point of comparison in the empirical results which will be discussed in subsequent chapters.

1.1 Establishing and Analysing a Large Scale Textual Corpus

The first step in a corpus based approach to natural language processing is the selection of the data which will provide the basis for our model. I've picked the English language portion of Wikipedia as my data source, a choice which is in accordance with a good deal of work done in the field. For instance, Gabrilovich and Markovitch (2007) and Collobert and Weston (2008), to name just a couple, use Wikipedia as their base data for training distributional semantic models designed to perform tasks similar to the ones explored in subsequent chapters, while **?**, Pennington, Socher, and Manning (2014), and Gutiérrez, Shutova, Marghetis, and Bergen (2016) use amalgamated corpora that include Wikipedia as a major component. Wikipedia provides a very large sample of highly regular language, meaning that we can expect a certain syntactic and semantic consistency as well as language which, if not always overtly literal, is likewise not typically abstruse or periphrasitc. This should supply a source of linguistic data in which, to revisit the central dogma of the distributional hypothesis, words which occur in a specific syntactic and lexical setting can be expected to be semantically similar.

In the case of my implementations, the November 2014 dump of English language

Wikipedia has been used.[1] A data cleaning process has been implemented, the first step of which is the chunking of the corpus into individual sentences. Next parenthetical phrases are removed from each sentence, as these can potentially skew co-occurrence data, and all other punctuation is subsequently removed. All characters are converted into lowercase to avoid words capitalised at the beginning of sentences, quotations, and other places being considered as unique types. Finally, the articles *a*, *an*, and *the* are removed as they can distort co-occurrence distance counts. The cleaned corpus contains nearly 1.1 billion word tokens, consisting of almost 7.5 million unique word types. The distribution of these types is predictably Zipfian: over 10 million occurrences of the top nine word types are observed, while the least frequent 4.27 million words – more than half of all types – only occur once. The top end of this distribution is populated by conjunctions, prepositions, and pronouns, while the bottom end is characterised by obscure place names, one-off abbreviations, unicode representing non-Latin alphabet spellings, and a good many spelling errors.

As is generally the case with data cleaning, these measures are prone to error: for instance, due to the removal of punctuation, the contraction *we're* will be considered identical to the word *were*. One of the strengths of the subspace projection technique that my methodology uses is its resilience to noise. So, for instance, misspellings will be categorised as highly anomalous co-occurrence dimensions and are therefore unlikely to be contextually selected – or, if they are regularly encountered enough to be contextually significant, there may well be useful information in the co-occurrence profile of such mistakes – and, at the other end of the spectrum, essentially ubiquitous words are unlikely to provide context specific information, so the ambiguity between *we're* and *were* is unlikely to be drawn into any of the subspaces actually projected by the model.

From the cleaned corpus, a model's vocabulary is defined as the top 200,000 most frequently occurring word types. This cut-off point is very close to the point where the total number of word tokens included – that is, occurrences of any word of any type

[1] Relatively recent Wikipedia dumps are available at `https://dumps.wikimedia.org/`.

– by selecting all instances of all vocabulary words equals the total number of word types – that is, unique word forms – excluded. Given the Zipfian distribution of word frequencies as observed throughout the corpus, this means that more than 95% of the co-occurrence data available from the corpus will be taken into account by the model, while the number of word-vectors used to express this data represents less than 5% of the potential vocabulary—a fairly efficient way of extrapolating statistics from the corpus. The selection of this as a cut-off point means that the least frequent words in the vocabulary occur 83 times throughout the corpus.

Having processed the corpus and established the target vocabulary, the next step of this methodology is to build up a based space of co-occurrence statistics. Here, following the example of the majority distributional semantic work, co-occurrence between a word w and another word c will be considered in terms of the number of other words between w and c. In the case of my methodology, and again in accord with the a great deal of work within the field, a statistic for word w in terms of its co-occurrence with c will be derived from the consideration of all the times that c is observed within k words of w, where k is one of the primary model parameters that will be considered in the experiments reported in later chapters of this thesis. Based on these co-occurrence events, a matrix M is defined, where rows consist of word-vectors, one for each of the 200,000 words in the vocabulary, and columns correspond to terms with which these vocabulary words co-occur. These column-wise co-occurrence dimensions include the words in the vocabulary as well as many, many words that are not in the vocabulary, to the extent that every word type in the corpus is considered as a candidate for co-occurrence. A *pointwise mutual information* metric gauging the unexpectedness associated with the co-occurrence of two words is calculated in terms of this equation:

$$M_{w,c} = \log_2 \left(\frac{f_{w,c} \times W}{f_w \times (f_c + a)} + 1 \right) \tag{1.1}$$

Here $f_{w,c}$ represents the total number of times that c is observed as co-occurring in a sentence within k words on either side of w, f_w is the independent frequency of occurrences of w, and f_c is likewise the overall frequency of c being observed as a co-occurrence term throughout the corpus. W is the overall occurrence of all words throughout the corpus–and it should be noted that, excluding the term a, the ratio in Equation 1.1 is equivalent to the joint probability of w and c co-occurring. The term a is a skewing constant used to prevent highly specific co-occurrences from dominating the analysis of a word's profile, set for the purposes of the work reported here at 10,000.[2] Finally, the entire ratio is skewed by 1 so that all values returned by the logarithm will be greater than 0, with a value of zero therefore indicating that two words have never been observed to co-occur with one another.

This last step of incrementing the ratio of frequencies in order to avoid values tending towards negative infinity in the case of very unlikely co-occurrences is again a departure from standard practice, where, in word counting models, a *positive pointwise mutual information* mechanism involving not skewing the ratio and instead treating any ratio of frequencies less than 1 – that is, any co-occurrence that is observed less often than the balance of the mean values for all occurrences of w and all co-occurrences with c – as being equivalent to zero, or no co-occurrence at all (Levy and Goldberg, 2014, have considered a more general variable ratio shifting parameter). The motivation for this more typical technique is again to avoid incorporating unnecessary and potentially confounding information into a model, but, again, in the case of my model, the dimensional selection process will tend to ignore such information, and at the same time, as will be seen, data regarding relatively unlikely co-occurrences can sometimes also be quite informative. Other areas for variation in deriving co-occurrence statistics include the nature of the co-occurrence window itself, where some researchers have taken weighted samples (**?**)r considered word order, and also the actual representation of tokens within

[2] Anecdotally, the first combination of input words analysed during an early stage of the development of this model that didn't use a smoothing constant was the phrase "musical creativity", and the very first dimension indicated by the analysis was labelled *gwiggins*—the email handle of one of my supervisors. Prof. Wiggins's deep connection with music and creativity meant that every instance of *gwiggins* occurring throughout Wikipedia was in the vicinity of both *musical* and *creativity*, and so the dimension was indicated by the combination of these terms, which makes sense, but it was still a bit eerie to have such a personally relevant result generated by a model based on such general data.

the corpus, where part-of-speech and dependency tagging (Padó and Lapata, 2007) have been applied to positive effect. Lapesa and Evert (2014) and Milajevs, Sadrzadeh, and Purver (2016) offer comparative overviews of the effects of parameter variations on the performance of distributional semantic techniques.

The net result of my methodology is a matrix of weighted co-occurrence statistics, where higher values indicate a high number of observations of word w co-occurring with word c relative to the overall independent frequencies of w and c. Values of zero indicate words which have never been observed to co-occur in the corpus, and, as most words never co-occur with one another, the matrix is highly sparse. The weighting scheme results in a kind of semi-normalisation of the matrix: infrequent words will tend to correspond to more sparse dimensions, but the non-zero values along these dimensions will by the same token tend to be higher due to the lower value of the word's frequency in the denominator. So far this technique sits comfortably within the scope of existing work in the field. It is what I propose to do with this base matrix that will begin to distinguish my methodology, and this next step in the process of projecting context sensitive spaces of word-vectors will be discussed in the following section.

1.2 Selecting Dimensions from a Sparse Co-Occurrence Matrix

Context has thus far remained a somewhat abstract concept in this thesis. In principle, the context in which conceptualisation occurs for a cognitive agent is its environment with all its affordances, linguistic and semantic but also more generally perceptual: in a word, the agent's *umwelt* (von Uexküll, 1957). In the world of physical entanglements, language presents itself with precisely the same open-ended opportunities for action as other modes of cognition—and, in the case of language, the action afforded is meaning making. In practice, however, for the purposes of my methodology, context will be defined lexically, as a word or set of words which are fed to a model, analysed in terms of their co-occurrence profiles, and then used to generate a subspace of conceptually relevant co-occurrence dimensions. The intuition behind this approach is the idea that there should be a set of words which collectively selects a set of dimensions that are conceptually relevant to some conceptual context, and the geometry of the word-vectors of my model vocabulary as projected into the subspace delineated by this set of dimensions should reveal the semantics of this context.

So, notwithstanding interesting work on multi-modal approaches to distributional semantics from, for instance, Hill and Korhonen (2014) and Bruni, Tran, and Baroni (2014), with regard to the present technical description, I will treat *context* as meaning some set of words T which have been selected for the purpose of performing some type of semantic evaluation and act as input to a context sensitive distributional semantic model. The exact mechanisms for specifying T will be discussed in subsequent chapters with regard to each of the individual experiments to be performed using my methodology; for now, I offer a general outline. Each component of T points to a word-vector in the matrix M described in the previous section, and the collection of word-vectors corresponding to T serve as the basis for an analysis leading to the projection of a context specific subspace S. I propose three basic techniques for generating these projections, with the model parameter d indicating the specified dimensionality of the subspace to be selected:

Joint A subspace of d dimensions with non-zero values for all elements of T and the highest mean PMI values across all elements of T is selected;

Indy The top $d/|T|$, where $|T|$ is the cardinality of T, dimensions are selected for each element of T regardless of their values for other elements of T, and then these dimensions are combined to form a subspace with dimensionality d;

Zipped The top dimensions for each element of T are selected as in the INDY technique, with the caveat that all selected dimensions must have non-zero values for all elements of T and no dimension is selected more than once.

These techniques are used for the purpose of analysis, and, once this analysis has been performed, the subset of dimensions returned is used to project the entire model

		guitar			*dulcimer*	
	dimension	PMI	normalised	dimension	PMI	normalised
HIGH	*mandolin*	8.30964	0.10719	*hammered*	13.97749	0.09354
	bass	8.08501	0.10429	*dulcimer*	12.73992	0.08526
	12-string	8.07679	0.10418	*autoharp*	11.50399	0.07699
	acoustic	7.99076	0.10308	*appalachian*	11.23224	0.07517
	banjo	7.96400	0.10057	*zither*	10.98302	0.07350
LOW	attacked	0.05222	0.00067	*him*	0.25698	0.00172
	report	0.04768	0.00062	*school*	0.25340	0.00170
	country	0.04418	0.00057	*would*	0.23825	0.00159
	champions	0.02644	0.00034	*into*	0.21336	0.00143
	regions	0.02538	0.00033	*there*	0.21320	0.00143

Table 1-A: The top five and bottom five dimensions by PMI value for the words *guitar* and *dulcimer*, out of all the dimensions with non-zero values for both words, with scores tabulated independently for each word.

vocabulary onto a d dimensional subspace. The JOINT technique requires the greatest finesse, as there is an element of cross-dimensional comparison at play. As such, for the purposes of this technique, the word-vectors selected by T are merged, dimensions with non-zero values for any of the word-vectors are discarded, and the resulting truncated word-vectors, each consisting of an equal number of non-zero dimensions, are normalised. This ensures that certain elements of T won't dominate the analysis: because the frequency of each word in T applies a deflationary pressure on the PMI values associated with the corresponding word-vectors, very infrequent words would be liable to dominate the analysis with the associated high PMI values in their profile. This effect is illustrated in Table 1-A, where PMI values for the top dimensions selected using the JOINT type subspace by the words *guitar*, which at 88,285 occurrences is ranked 1541 in frequency, are compared with those for the word *dulcimer*, which occurs 516 times and is ranked 62,313 (the base model here was constructed using a 5x5 word co-occurrence window). Among the dimensions with non-zero values for both words, normalisation brings the respective co-occurrence profiles more in line with one another, facilitating the selection of a subspace which is jointly characteristic of the input terms.

In the cases of the INDY and ZIPPED techniques, the selectional process is more straightforward, since mean values between word-vectors are not being considered. Where the JOINT technique is intended to discover subspaces that represent an amalgamation of the input terms, the INDY technique is expected to produce a subspace where individual conceptual characteristics of the input terms, captured as collections of co-occurrence dimensions, are distilled into distinct geometric regions. The ZIPPED technique might be seen as something of a hybrid of the JOINT and INDY techniques, since it used the INDY approach to make selections from the intermediary space of non-zero dimensions available to the JOINT technique. In each instance, these techniques are formulated to return a set of dimensions which, with varying degrees of cohesion, delineate a space that is in some sense salient to the contextual terms T serving as the basis for the analysis. In all cases, these techniques are used for the purpose of analysis, and, once this analysis has been performed, the subset of dimensions returned is used to project the entire model vocabulary onto a d dimensional subspace.

In order to offer a sense of what's happening with these dimensions selection techniques, a preliminary and intuitively motivated case study of dimension selection is outlined in Table 1-B, again derived from a base space generated through observations made within a 5x5 word co-occurrence window over the course of the corpus. The top dimensions selected by each technique are presented for two different three term sets of input words: *lion*, *tiger*, and *bear*, on the one hand, which are taken to represent in their union exemplars of wild animals, and on the other hand *dog*, *hamster*, and *goldfish*, which are prototypical pets. The dimensions selected by the JOINT technique in response to the WILD ANIMAL type input include the names of other wild animals, as well as *paw*, a component of many wild animals, *mauled*, an activity performed by wild animals, and, interestingly, *mascot*, presumably because many sports teams take these types of animals as their mascot: while this connection may not be immediately intuitive, it seems likely that this word would probably select for other wild animals in terms of its co-occurrence profile. The dimensions returned by the INDY technique, on the other hand, are, as expected, more independently characteristic of each of the input terms, with culturally referential words like *cowardly* (presumably from many mentions of the

lion, tiger, bear			*dog, hamster, goldfish*		
JOINT	INDY	ZIPPED	JOINT	INDY	ZIPPED
leopard	cowardly	cowardly	pet	sled	dog
cub	crouching	sumatran	hamster	hamster	hamster
hyena	localities	grizzly	goldfish	goldfish	goldfish
sloth	rampant	tamer	hamsters	hound	pet
lion	sumatran	leopard	domesticated	djungarian	hamsters
mascot	grizzly	teddy	breed	koi	fancy
paw	wardrobe	tamarin	fancy	nassariidae	breed
tiger	leopard	tiger	pets	ovary	siberian
rhinoceros	stearns	polar	bred	carp	domesticated
mauled	teddy	passant	robotic	ednas	cat

Table 1-B: The top 10 dimensions returned using three different dimensional selection techniques, featuring one set of input terms collectively referring to wild animals and another set collectively referring to pets.

Cowardly Lion character from *The Wizard of Oz*) and *crouching* (indicating the popular Chinese movie *Crouching Tiger, Hidden Dragon*), as well as other species-specific terms such as *sumatran* and *grizzly.* Notably, the term *stearns* pops up here, certainly because of prolific references on Wikipedia to the defunct investment bank Bear Stearns, illustrating ways in which the INDY technique might allow for dimensions indicative of underlying polysemy.

Similar effects are observed in response to the PET type input. The word *pet*, two of the three input terms themselves, and the names of other types of pets appear in the output from the JOINT technique, as well as descriptive terms such as *domesticated*, *breed*, and, amusingly but not irrelevantly, *robotic*, presumably because of the phenomenon of robotic pets, which has its own page on Wikipedia. The INDY technique, on the other hand, returns some very term specific dimensions, again indicating a degree of ambiguity, such as *djungarian* (a breed of hamster popular as a house pet), *nassariidae* (in fact a species of snail, known colloquial as the *dog whelk*), and *ednas* (Edna's Goldfish was a short-lived American punk rock band). In the cases of both PETS and WILD ANIMALS, the dimensions returned by the ZIPPED technique represent something of an intermediary between the two other techniques, tending to include some of the terms generated using the JOINT technique but also some more word-specific terms. The actual geometry of these spaces will be discussed generally in the next section, and will be explored in detail in relation to specific semantic applications in subsequent chapters.

A very broadly similar approach to distributional semantics has been proposed by Polajnar and Clark (2014), who describe a *context selection* methodology for generating word-vectors, involving building a base space of co-occurrence statistics and then transforming this space by preserving only the highest values for each word-vector up to some parametrically determined cut-off point, setting all other values to zero. Setting the cut-off point relatively stringently – generating a base space of more sparse word-vectors, followed by various dimension reduction techniques – led to improvements in results on both word similarity and compositionality tests. This suggests that allowing word-vectors to shed some of their more obscure co-occurrence statistics leads to a more sharply defined semantic space, and indeed there may be an element of disambiguation at play here, as well, with vectors dropping some of the numbers associated with less frequent alternate word senses.

In the end, though, the method described by Polajnar and Clark results in a space which, while the information contained in the representation of a particular word is to a certain extent focused on the most typical co-occurrence features of that word, is still fundamentally general and static. To the extent that any contextualisation takes place here, it happens *a priori* and is cemented into a fixed spatial relationship between word-vectors. This is anathema to the theoretically grounding of my methodology, which holds that conceptual relationships arise situationally, and that semantic representations should therefore likewise come about in an *ad hoc* way. The novelty, and, I will argue, the power of my approach lies in its capacity to generate bespoke subspaces in reaction to semantic input as it emerges, and the expectation is that these subspaces will have a likewise context specific geometry which can be explored in order to discover situationally significant relationships between the projected semantic representations. The next section will begin to examine how these geometries might look.

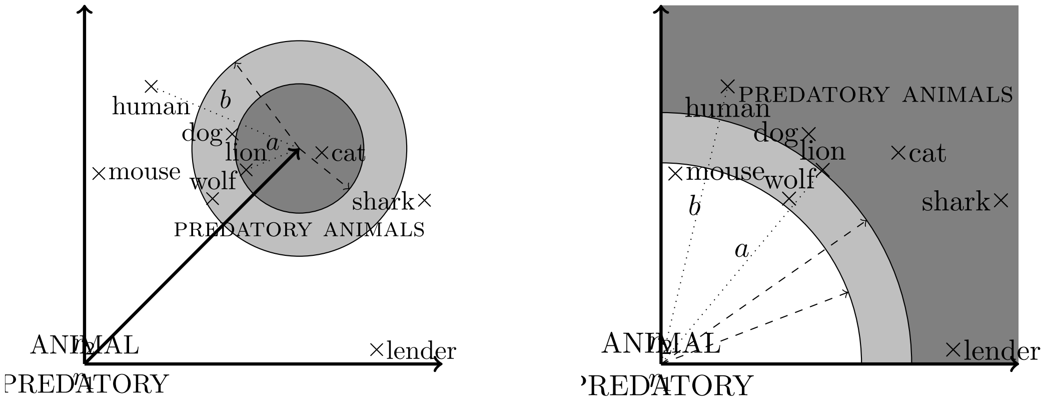
1.3 Exploring the Geometry of a Context Specific Subspace

Before delving into the question of the types of geometries my method might be expected to generate, I would like to raise a point regarding the typical application of the term *geometry* to vector space models of distributional semantics in the first place. Widdows (2004) makes an enthusiastic and compelling case for the representational power of geometry, while Clark (2015) has pointed out that treating words as geometric features endows lexical representations with "significant internal structure" (p. 509) which can be applied towards modelling the meaning making compositionality of language. Baroni, Bernardi, and Zamparelli (2014) go so far as to suggest that their distributional semantic model effectively instantiates the abstract principles of Frege's work on the logic of natural languages (Dummett, 1981) in a geometric mode. These are powerful points touching on the essence of semiotics, and the idea that representations that map from data to interpretable features in a space are core to my own methodology.

The point I would like to make now, though, is that there are different degrees of geometry that can be in principle accessed in a vector space of real valued dimensions. The great majority of approaches surveyed here, taken to be representative of the historical and ongoing trend in the field, present models consisting of spaces of normalised word-vectors, in which there is a monotonic correlation between the distance and the angle between two word-vectors. In the case of models built using a principal component analysis, this is because when eigenvectors are used as dimensions of maximal variance, there is no meaningful interpretation of sign along these dimensions; in fact, mean values along a dimension will tend towards zero and the signs of values along any dimension discovered through a singular value decomposition can be reversed without any degradation of the information available from the analysis (Abdi and Williams, 2010). So, while Euclidean distance is strictly meaningful in such a dimensional reduction, there is no sense of a centre of the space other than the centre of gravity of the data as projected onto the selected number of eigenvectors, and cosine similarity is in practice the measure used to determine the similarity between two word-vectors. And in the case of models built using neural networks, there is no meaningful interpretation of dimensions to begin with, so the resulting space is a *de facto* hypersphere of word vectors that are only relative in terms of their relationship to one another, not their relationship to any objective features of the space.

In the case of my methodology, however, precise values along dimensions, and, correspondingly, overall Euclidean distances are significant: because base dimensions are preserved in the spaces projected through any of the dimension selection techniques described above, the actual position of word-vectors in space, not just their relative situations on the surface of a normal hypersphere, are significant, with a number of potentially desirable effects. The first effect to note is that in my subspaces distance from the origin is expected to be a meaningful feature. In a subspace of contextually selected dimensions, word-vectors with strong co-occurrence tendencies for that set of dimensions should have high PMI values across all dimensions, and so a relatively high norm of a word-vector is anticipated to correspond to semantic saliency within that context.

The second effect is that there is a notion of centre and periphery in my subspaces. Since all values are positive, a word-vector with high scores across all or most dimensions in a subspace will be far from the origin and in the central region of the space. A further consequence of the positivity of these subspaces is that word-vectors with mainly low or null PMI values will be far from the centre, so in the end two word-vectors may be both close to the centre of a subspace, or at the periphery of a subspace but close to one another, or at the periphery and far from each other, at two different edges of the positively valued space, and each of these situations can be predicted to have a particular semantic interpretation. The third effect, which follows from the first two points, is that a subspace can be characterised in terms of a set of key points based on an analysis of the collective profiles of the dimensions delineating the subspace, by which I mean some



(a) Word-vectors measured by proximity to a central point.

(b) Word-vectors meansured in terms of distance from origin.

Figure 1.1: A sampling of data from the space described in Section **??** projected into a drastically reduced (from approximately 7.5 million dimensions to 2 dimensions) subspace defined in terms of the co-occurrence features *animal* and *predatory*. Terms which are animal but not predatory (*mouse*), or predatory but not animal (*lender*) fall towards the edges.

straightforward assessments of the statistical distribution of each dimension involved.

1.3.1 Two Measures for Probing a Subspace

In order to take a first pass at examining these robustly Euclidean features of my contextualised subspaces, I propose two geometric measures for exploring the conceptual geometry of a subspace, illustrated in Figure 1.1. The first is a distance metric, which defines a central point in a subspace and then considers the relationship of words to the semantic context of the subspace in terms of the distance of the corresponding word-vectors from this central point. In practice, the central point will be defined as the mean point between the input word-vectors used to generate the subspace, but for the purposes of Figure 1.1a, a point along the line extending from the origin through the centre of the two dimensional space has been chosen. In this subspace featuring two hand picked co-occurrence dimensions selected from a base model built from a 5x5 word co-occurrence window traversal of Wikipedia, word-vectors relatively closely associated with the concept PREDATORY ANIMAL turn up near this central point.[3] So, for instance, cats (certainly in their taxonomical sense), more specifically lions, dogs, and, again more specifically, wolves all fall close to the central point, while sharks (certainly predators, and also animals, but perhaps less prototypically so), mice, humans, and lenders are more distant.

The second measure deployed here will be to analyse the norms of the word-vectors projected into the contextualised subspace, with my hypothesis being that word-vectors that are relatively far from the origin will be correspondingly relevant to the conceptual context from which the subspace has been generated. This prediction does not entirely play out in the subspace depicted in Figure 1.1b, where words like *human* and *lender* are about as far from the origin as *cat* and *shark*, and have higher norms than more prototypical denotations such as *lion* and *wolf.* As will be seen in subsequent results, beginning here and extending into the experiments described in the next chapter, in higher dimensional subspaces selected using the techniques outlined above, norm does prove to be a predictive measure of semantic relevance. Here again, the preponderance of co-occurrence statistics associated with a word over the course of a set of dimensions gives a higher dimensional subspace an advantage: if the selected dimensions are appropriately aligned, there will be a tendency for those word-vectors with some consistency of co-occurrence across all dimensions to extend towards the central fringe of the space, while those with inconsistent co-occurrence profiles will move towards the edges while remaining closer to the origin.

In the cases of both the distance from mean and norm measures, a threshold could, in principle, be established in order to determine a cut-off point for conceptual membership, either in terms of an absolute geometric measure – a radius from either the central point or the origin – or in terms of a set of nearest neighbours. This move would begin to move these subspaces towards Gärdenfors's (2000) notion of a region within a conceptual

[3] Here it happens to be the case that choosing dimensions which actually nominate a concept likewise delineate a space where, at least in terms of the restricted vocabulary evoked in Figure 1.1, conceptual membership plays out in a geometrically predictable way, but I will not generally presume this to be the case.

lion, tiger, bear					
JOINT			INDY		
norm	distance	angle	norm	distance	angel
leopard	cat	and	leopard	wild	and
langur	wild	like	dhole	cat	as
hyena	wolf	also	hyena	giant	which
dhole	elephant	as	rhinoceros	elephant	like
boar	animals	such	leopards	lions	also
tapir	giant	well	tapir	wolf	be
macaque	animal	including	passant	animals	more
chital	bears	include	langur	tigers	including
civet	dog	from	sumatran	cats	been
sloth	panther	which	gules	golden	one

Table 1-C: The top word-vectors in subspaces selected by input terms characteristic of WILD ANIMALS, for the JOINT and INDY dimension selection techniques, measured in terms of top norms within each subspace, word-vectors closest to the mean point between the input word-vectors, and also the smallest angle with this mean vector regardless of actual position in the subspace.

space, particularly in the case of the distance based metric illustrated in Figure 1.1a: here a clear sense of convexity as a criterion for a conceptual region exists, and likewise of betweeness as an indicator of conceptual inclusion. Importantly, though, these spaces as they stand lack the dimensional interpretability that characterises Gärdenfors's spaces, in that it is not possible to say that there is a dimension of size, or strength, or ferocity, or so forth along which a boundary for inclusion in the concept of PREDATORY ANIMAL can be identified.

Examples of the tendencies of both norms and relative distances are explored in Table 1-C and Table 1-D, where, as with the examples offered earlier in this chapter, input terms denoting things exemplary of the respective concepts WILD ANIMALS and PETS are used to generate subspaces, in this case using both the JOINT and INDY dimension selection techniques, once again using a base space built using a 5x5 word co-occurrence window. In these cases, the top 200 dimensions derived using each technique have been used to project subspaces, and then within those subspaces, the top ten word-vectors based on their norm and their distance from the mean point between the input word-vectors are reported. In addition to the two geometric measures described above, as a

dog, hamster, goldfish					
JOINT			INDY		
norm	distance	angle	norm	distance	angle
hamsters	cat	and	dogs	cat	also
gerbils	pet	also	hamsters	giant	as
rabbits	monkey	as	sheepdog	animal	in
chinchillas	pig	of	terrier	wild	which
pet	rabbit	in	canine	animals	and
ferrets	rat	such	kennel	like	like
pigs	animal	well	akc	rabbit	is
rats	dogs	-	spaniel	include	called
pets	giant	called	poodle	pig	of
chickens	cats	which	jerboa	cats	has

Table 1-D: The top word-vectors in subspace, as in Table 1-C but selected by input terms characteristic of PETS.

point of comparison, I also present results using an angular measure, where the word-vectors with the highest cosine similarity with the vector of the mean point between the input word-vectors. This is offered as an approximation of what would be a typical approach in a standard static distributional model, to demonstrate why this measure doesn't work for the context sensitive spaces built using my methodology and also as a mechanism for further exploration of what's happening in these subspaces.

Notably, in the case of the norm measure, word-vectors that are exemplary of the conceptual category suggested by the intersection of the input terms seem to rise to the top of the subspace, so to speak: for both dimension selection techniques for the WILD ANIMAL type inputs, a list of wild animals, some rather exotic, are returned. A similar outcome is observed for the norm measure in the case of the pet inputs, with some admittedly disputable admissions such as *rats* coming up in the JOINT output; jerboas, which are indicated in the INDY output, are apparently a somewhat popular pet, and *akc* presumably refers to the American Kennel Club, so, not a pet, but an institution related to pet keeping. An interesting side effect of the INDY technique in particular is that it returns list including names of various dog breeds. It would seem that the co-occurrence dimensions of the word-vectors for *hamster* and *goldfish* are characteristic enough of these more specialised words relating to particular types of pets that the corresponding word-vectors are pushed towards the outer fringe of the subspace. It's also interesting that *passant* and *gules*, terms associated with the depiction of animals in heraldry, have high norms in the INDY space in particular—of course all three of the input terms here are denotations of animals typical of heraldic devices, so it is not particularly surprising that some of their independently strong co-occurrence features combine to select for these word-vectors.

The distance measure returns roughly similar results, including a number of denotations of appropriate animals. Here it is interesting to observe that other semantic types – in particular, adjectives in addition to nouns – begin to creep into the output: *wild*, *giant*, and *golden* are returned in the JOINT and INDY subspaces for the *wild animals* input, and *giat* again comes up in response to the PETS input, along with, perplexingly, the verb *include*. It makes sense that the region near the mean point between the input vectors, where consistently high but perhaps not absolutely maximal PMI scores across these contextually characteristic dimensions are to be found, feature some of the descriptors and predicates associated with the concept being modelled, while the region at the outer fringe of the space, where the words with the highest overall PMI values across the dimensions of the subspace, would be pointed denotations of instances of the concepts in question. The word-vectors corresponding to some of the more esoteric animals in particular are likely to have high co-occurrence frequencies with the same dimensions selected by the combination of the input terms relative to low independent frequencies precisely because of their rareness.

Turning to the angular results, where words that are closest to the line extending through the mean point are returned, a sharp contrast to the other two geometric measures is observed. Here, very generic words which serve as the structural components of language, contributing little in terms of specific meaning but crucial to the functional cohesion of an utterance, are found in abundance. This is completely logical: these types of words are liable to have a very consistent, albeit relatively low, profile of PMI scores across all dimensions in a subspace, since they are likely to have a high frequency of co-occurrences with any given word mitigated by a correspondingly high independent frequency across the corpus influencing the denominator of the PMI calculation. The result is a word-vector populated by relatively low but also relatively consistent PMI values, situated not far from the origin and also very close to the centre of the subspace. This phenomenon highlights the discrepancy between the Euclidean, positively valued subspaces generated by my context sensitive methodology and the normalised, hyperspherical spaces built by conventional static distributional semantic models. Because my subspaces have a sense of centre and periphery, as well as a sense of distance from the origin, it is possible to make both semantic and functional predictions about the types of words that will be found in different regions of a subspace, and accordingly to predict where to look – and where not to look – to discover geometries mapping to desired conceptual properties.

1.3.2 Replete Geometric Analysis

I will now propose a general method for a replete geometric analysis of a contextually projected subspace, based on the position of word-vectors in a space as well as the relationship between those word-vectors and points based on a more general analysis of the dimensions delineating the subspace. For the purposes of explicating this method, I will presume a subspace projected from an analysis of two input word-vectors A and B using one of the dimension selection techniques described earlier in this chapter. The presumption is that these word vectors are to be analysed in terms of their semantic relationship; the precise nature of the relationship being analysed could be more or less anything, and in the next two chapters this method will be applied to the assessment of lexical similarity, relatedness, metaphor, and metonymy. The objective of this analytic method will be first to test the hypothesis that the geometry of contextually projected subspaces should be semantically informative, and second to compare the aspects of the geometry that are most informative for different semantic phenomena.

Figure 1.2 illustrates a generic three dimensional subspace. Points A and B are the

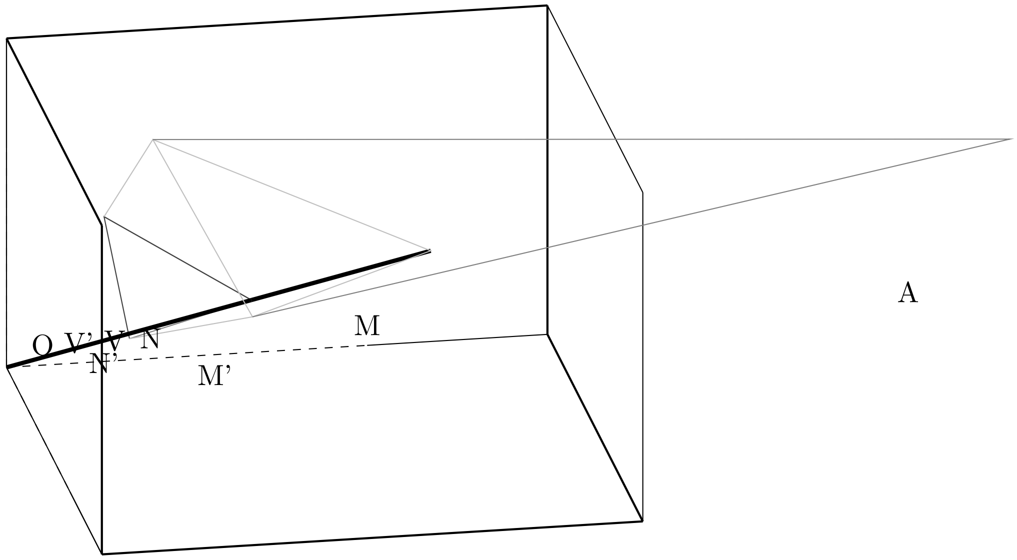


Figure 1.2: The geometric features of a subspace contextually projected based on an analysis of two input word-vectors.

two word-vectors that have been used to select the dimensions which define this subspace, and are likewise the word-vectors which will be analysed through the geometry of the subspace. In addition to these two points explicitly defined in terms of the values of projected word-vectors, two points are established based on an overall analysis of the dimensionality of the subspace: the *mean point* M and the *maximal point* X. M is defined as the vector of all the mean values for all the dimensions J delineating the subspace, so, if the dimensionality of J is d, M can be defined formally as follows:

$$M = \{\mu(J_1),\ \mu(J_2)...\ \mu(J_d)\} \tag{1.2}$$

And likewise, X can be expressed in terms of an equation:

$$X = \{max(J_1),\ max(J_2)...\ max(J_d)\} \tag{1.3}$$

Finally, a generic central point C, a vector with all dimensions set to the same value, is defined. The universal value chosen to define the dimensions of this vector is the mean value of the mean point M, so, formally, this point is the vector of that mean value repeated d times:

$$C = \{\mu(M),\ \mu(M)...\mu(M)\} \tag{1.4}$$

In the analysis of the semantic relationship between A and B in a given projection, these three vectors will be used as anchor points to establish the situation of A and B relative to the subspace overall: where C is an objectively central point in the subspace, M is in a sense central to a subspace relative to its particular dimensional constitution, and X is similarly indicative of the outermost possible extent of a particular subspace. The underlying point here is that, due to the frequentist components of the information theoretic co-occurrence statistics used to build the base space described here, different dimensions have different distributional profiles. To demonstrate this point, Table 1-E presents the mean value and standard deviations for the distribution of mean and maximum points from the top 20,000[4] most frequent co-occurrence dimensions, as well as the top five and bottom five values for each of these statistics for illustrative purposes.

The

1.4 Comparing to Alternative Approaches

1.4.1 Static Interpretations of the Base Space

1.4.2 A Model Trained Using a Neural Network

[4] less frequent dimensions tend to have higher PMI values overall, and also tend to be products of co-occurrences observed in quite obscure passages of the base corpus—it's worth recalling that a little more than half of the co-occurrence dimensions are observed only once.

	MEAN	MAX
TOP	*sofla:* 6.984	*nico:* 15.690
	olya: 6.326	*yeah:* 15.610
	non-families: 6.035	*superfamily:* 15.598
	gmina: 5.364	*eel:* 15.483
	crambidae: 5.485	*kermanshah:* 15.455
BOTTOM	*it:* 0.748	*he:* 3.903
	they: 0.812	*in:* 3.449
	you: 0.804	*of:* 3.379
	this: 0.789	*to:* 3.120
	he: 0.719	*and:* 2.993
mean	2.312	11.066
std	0.396	1.607

Table 1-E: Dimensional profiles in terms of mean and maximum PMI values along dimensions, including mean values and standard deviation as well as the top five and bottom five dimensions for each statistic.

References

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

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

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

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

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

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

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

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

Gutiérrez, E. D., E. Shutova, T. Marghetis, and B. K. Bergen (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 A. Korhonen (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*, pp. 255–265.

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

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

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

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

Polajnar, T. and S. Clark (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*, pp. 230–238.

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

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