Journal of Artificial General Intelligence 6(1) 55-86, 2015
DOI: 10.1515/jagi-2015-0004

Submitted 2015-05-24
Accepted 2015-11-19

From Distributional Semantics to Conceptual Spaces: A Novel Computational Method for Concept Creation

Stephen McGregor[⋆] S.E.MCGREGOR@QMUL.AC.UK
Kat Agres[⋆] KATHLEEN.AGRES@QMUL.AC.UK
Matthew Purver M.PURVER@QMUL.AC.UK
Geraint A. Wiggins GERAINT.WIGGINS@QMUL.AC.UK
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
[⋆] ***Primary authors contributing equally to this work.***

Editor: Tarek R. Besold, Kai-Uwe Kühnberger, Tony Veale

Abstract

We investigate the relationship between lexical spaces and contextually-defined conceptual spaces, offering applications to creative concept discovery. We define a computational method for discovering members of concepts based on semantic spaces: starting with a standard distributional model derived from corpus co-occurrence statistics, we dynamically select characteristic dimensions associated with seed terms, and thus a subspace of terms defining the related concept. This approach performs as well as, and in some cases better than, leading distributional semantic models on a WordNet-based concept discovery task, while also providing a model of concepts as convex regions within a space with interpretable dimensions. In particular, it performs well on more specific, contextualized concepts; to investigate this we therefore move beyond WordNet to a set of human empirical studies, in which we compare output against human responses on a membership task for novel concepts. Finally, a separate panel of judges rate both model output and human responses, showing similar ratings in many cases, and some commonalities and divergences which reveal interesting issues for computational concept discovery.

Keywords: distributional semantics, conceptual spaces, computational creativity, concept discovery, behavioural validation

1. Introduction

This paper presents a computational model for the discovery of concepts which attempts to bridge the gap between standard lexical distributional semantics and conceptual spaces. Beginning with a general vector space model of word meaning derived from co-occurrence statistics, we use input query terms to select a contextualized sub-space corresponding to a concept, and discover members of the concept as vectors within that space. Our general objective is to investigate whether a model which situates words in a spatial relationship to one another can be mapped to likewise spatially calibrated models of concepts. Such a model may be thought of as a method for modelling and discovering meaningful conceptual relationships, a topic of interest to general artificial intelligence. The power of our approach lies in its ability to draw this connection between an arrangement of terms in a lexical space and a representation in a cognitive space: clusterings of words are discovered

in a subspace informed by an input query, and this subspace suggests the structure, in terms of constituency, of a parallel conceptual region.

Vector space models of distributional semantics are currently a popular approach for quantifying similarity in computational linguistics, but many contemporary studies need grounding and external validation (Clark, 2015). Much of the work in this area compares model performance to semantic databases, but does not directly relate results to the cognitive performance of humans, or uses very restricted targets, such as word similarity judgments, rather than higher-level *concepts*. On the other hand, cognitive scientists have seen vector space models as a suitable way to model concepts, capturing notions such as conceptual similarity, degree judgements, prototypical membership and the effects of context on these (Gärdenfors, 2000). Our aim here is to investigate the connection between the two, and in the process to elucidate how humans conceptualise creativity; we therefore examine our model and compare to state-of-the-art approaches on tasks related to concept discovery, and include evaluation against human responses.

The model we propose is based on a standard approach to distributional lexical semantics, but differs from standard models in a few important aspects. First, we propose a contextualised method of dimension selection, rather than the standard approach of general dimensional reduction: we choose significant dimensions of the model based on the seed terms given (taken to name or define the target concept), and use these to outline a concept-specific subspace. By exploring this space we can suggest members of the concept, and we show that this provides accuracy at least as good as state-of-the-art distributional semantic models on a WordNet-based task, while also providing many of the properties of conceptual space models: the subspace is convex and relatively low-dimensional, and defined by interpretable characteristic dimensions. In particular, our approach performs better on concepts which are neither very general, common, abstract concepts nor tightly-defined scientific concepts, both of which are modelled well by standard approaches. We then move beyond the limitations of WordNet's taxonomy to investigate these more novel and unusual concepts via comparisons to human judgements.

The discovery and modeling of novel concepts is not a sufficient property for a model of computational creativity, but it is a necessary one. By seeking to implement a low level approach to the delineation of conceptual regions based on the geometry of a distributed semantic space, and one which views concepts as momentary and pragmatic phenomena which can emerge in a context without predefinition or preformulation, we hope to contribute to the understanding of methods which can perform this creative task. Furthermore, by investigating the use and limitations of lexical ontologies such as WordNet, and the supplementation thereof via human judgment studies, we hope to contribute to the understanding of suitable methods for evaluation of creative behaviour. Although many computational and AI systems aim to display creative behaviour or produce creative artefacts, the evaluation of computational creativity remains challenging and controversial, and in some cases, the issue of evaluation is not even explicitly addressed. This secondary aspect of the work, the potential for meta-analysis inherent in the question of whether our model's output will be useful for guiding an evaluative discussion of creative work elsewhere, is intended to give the work its own pragmatic grounding, in that this suggests a practical application for the creative output described in the following pages.

The organization of the paper is as follows: first we offer an overview of geometric models for concepts and words, together with the relation between them, positing that a crucial correspondence between the cognitive and linguistic domains can be found by situating them both spatially. This is followed by a general description of our methodology, which involves building a distributed

semantic space and then projecting subspaces from this base space informed by the conceptually contextual information contained in an input query. Based on this methodology, we then evaluate our approach and compare it to existing general lexical approaches on a task of discovering conceptual members, deriving models from a large scale corpus (the English language edition of Wikipedia), and evaluating against a set of classes in WordNet. We go on to investigate more creative cases via two empirical studies comparing human conceptualisations with our model's output: first by comparing automatic outputs to human-generated equivalents, and second by asking human judges to blindly compare our model's output to human generated terms. For the majority of concepts, our model performs comparably to humans. We conclude with a summary of our results and a brief consideration of possible avenues for future work.

2. Concepts and Creativity

As Barsalou (1993) has pointed out, when concepts are discussed in formally theoretical contexts, they are almost always construed in terms of words: in fact, it is almost impossible to imagine performing an analysis of conceptualisation without immediately resorting to words. Yet concepts clearly do not supervene on the words that can be used to denote and perhaps to outline them; cognitive content is seemingly something other than language, and certainly something other than the day-to-day language documented in a lexicon. On the other hand, it is even less clear how exactly a computer could be used to model conceptualisation, and here we must surely resort to some sort of commerce in linguistic symbols. Words are, as such, the sticking point between abstract qualitative conceptual processes and the likewise abstract but also essentially quantifiable computational operations involved in modeling conceptualisation.

There is something not true to life, though, in constructing a conceptual model as a look-up table, however comprehensive, merely associating words with rules. Rather, concepts seem to come about in the process of a cognitive agent's dynamic interaction with an environment, and in this there is something fundamentally creative: the ability to react to an unpredictable world necessitates the ongoing production of novel, meaningful information structures. So when we talk about conceptual creativity, what we are referring to is an agent's ability to form useful representations of a situation in an immediate and context sensitive way. The question of modeling such a process then becomes a question of the nature of the representations themselves, and in particular a matter of how they achieve their contextual dynamism.

The final piece of the picture to be painted here is a theory of conceptualisation as something that unfolds in a fundamentally spatial, geometric milieu—it is in this correspondence between concepts and space that the answer to the problem of representational structure is to be found. This section addresses the theoretical problem of how world relevant, informationally productive representations can be modeled, beginning with the question of the relationship between concepts in the world and the geometric capacities of symbol manipulating machines.

2.1 Conceptual Creativity and Computers

Koestler (1964) has proposed that creativity can be understood in terms of a *bisociative act*, by which some new meshing of previously disparate patterns of conceptual schemata results in the discovery of a new perspective on the world. Koestler characterises patterns of thought and behaviour in terms of *matrices*, and he likewise describes creativity in terms of the discovery, through *bisociation*, of "new, auxiliary matrices" which allows the creator to overcome some

obstacle in the world. Extending this idea to span conceptual domains including all varieties of language use as well as non-propositional conceptualisation, Fauconnier and Turner (1998) have presented a theory of *conceptual blending* which holds that even the most quotidian of concepts are formed through the online interaction of *mental spaces*, "small conceptual packets constructed as we think and talk, for purposes of local understanding and action," (p. 137). Critically, these spaces are characterised by dynamics which determine the way that they can interact in the course of mappings projected across diffuse spaces.

These constructs which construe creativity in terms of discovery and spatial dynamism fit nicely with the classic approach to creative cognition of Boden (1990), who describes creativity in terms of state spaces that can be explored and transformed. When it comes to computers modeling the creative development of concepts, it is important that emphasis be placed on the mechanism by which the conceptualisation is performed—so Wiggins (2006) describes a framework for assessing computational creativity in terms of *creative behaviour*, with the focus being here on the way in which the system operates rather than merely the output of that operation. Implicit in this operational sensitivity is an awareness of what Boden has described as *P*- (personal, local) versus *H*- (historical, global) *creativity*, and likewise, in the terminology of Ritchie (2007), the *inspiring set* which serves as the foundation for a creative agent's activity, with the agent attempting to extend a base set of creative artefacts without simply imitating them.

In the case of the operation of the model which will be described here, which seeks to map concepts from clusters derived from operations on geometric word representations, the fact that, for instance, the concept DOG is more or less paradigmatically similar to CAT may be inherent in the structure of the underlying data, which is to say, in the relationship between words in a large scale corpus that discusses things like cats and dogs. That CAT is sometimes akin to PANTHER, or indeed in certain circumstances to JAZZ MUSICIAN or perhaps to other less likely things is a somewhat more interesting distinction to make. It is ultimately the discovery both of the classifications, which may be to some degree inherent in the underlying data, and of the operational, representational context in which these classification qualify, something that is fundamentally an aspect of the system's own representational dynamics, that draws out the creativity being modeled through a computational approach to linguistics. In our conceptually contextual language model, the things that are being associated are linguistic representations – words, but words cast as dynamic structures – and the process of association involves the ongoing development of previously uncharted constructions in an astronomically immense state space of projections.

The thrust of the argument here is that creative conceptualisation is something more than just an itemisation, more than just a rearranging of things into predefined categories. A creative conceptualisation involves the creation of a new way of associating things, and it is both the associations themselves and the novel criteria for making the associations which are submitted as a creative event. Creativity happens in the course of an agent's entanglements with the world: in an unpredictable environment, flexibility and immediacy are paramount, and so the cognitive state of an agent must be tightly coupled with the environment. The process by which such an agent achieves goals and indeed survives must involve something more than merely indexing reactions from a list of possible scenarios, since, to the agent (as opposed to its programmer), such a list is not available: novel scenarios that arise must be understood, and their representations synthesised anew. Thus, the kind of conceptual creativity that we are, in a broad sense, trying to describe and demonstrate is very much at the core of cognition in general. In order to model this phenomenon of

mind and environment computationally, it is necessary to build representations which interact with both input and each other dynamically.

2.2 Conceptual Spaces

Gärdenfors (2000) has developed a geometrical theory of *conceptual spaces* which focuses on the dimensionality of the spaces, and the way that a space's dimensions can correspond to perception of phenomena in the world. A crucial characteristic of such a space is that the regions of the space, which might be construed as conceptual entities, can be seen as interacting with one another by virtue of the lower level relationships between their defining dimensions. Among other things, Gärdenfors' model presents a basis for resolving the difficult relationship between low level stimuli in an environment, which become dimensions in conceptual spaces, and the higher level interactions of representations apparent in cognitive processes.

Gärdenfors sees concepts as corresponding to regions in such spaces, with individual entities which belong to those concepts (or events etc.) as points within those regions. This provides a view naturally suited to modelling not only membership judgements in this way, but similarity judgements (via distances between points), the existence of prototypical members of concepts (as more central points within regions) and degree judgements (via distances from central points). Furthermore, Gärdenfors takes a critical property of *natural* concepts that they be representable as *convex* regions: any point that lies *between* two members of a region — given the notion of *betweenness* defined by the conceptual space and its dimensions — must also be a member of that region. This plays into the intuition that there cannot be gaps in the dimensional substrate of the conceptual manifold: a color that, in terms of brightness, lies somewhere between light red and dark red is still a shade of red. Furthermore, Gärdenfors presents a notion of *salience*, a factor that mediates the significance of certain aspects in the interactions between concepts: in the course of conceptual meshing, the interactive dimensions are weighted with regard to the significance of their role in the entanglement.

As will be described in Section 4, the model proposed here is grounded in the theoretical stance that concepts and conceptualisation can be understood in terms of geometry. Like Gärdenfors, we seek to develop a model of conceptualisation which is inherently spatial, and one which is able to develop concepts in a contextually sensitive way by virtue of the construction of dynamic linguistic representations which continuously interact with input in an unfolding process of ongoing conceptualisation. To this end, we are modeling creative conceptualisation, because our methodology facilitates the ongoing reaction to information as it arises in an unpredictable environment. Our model, however, operates in the rarefied but readily computable domain of corpus linguistics: where Gärdenfors describes a profusion of dimensions, both continuous and discrete, spanning colours, sensations, size, shape, time, physiognomy, society, and various other things both concrete and abstract, our model is confined to the narrow world of words as linguistic symbols and the way they relate to one another across the breadth of a textual corpus. Nonetheless, in the modern tradition of distributional semantics (see Section 3), we present a methodology for using the information inherent in a large body of text to define spatial regions relating to concepts, where both the relationships between positions and the dimensions describing the positions themselves are loaded with information. These regions we define are convex, and are defined by giving particular weight to contextually important dimensions, following Gärdenfors' approach to salience as well as

the dynamics described by Fauconnier and Turner (1998), desiderata fulfilled by the model's ability to contextually adjust the influence of dimensions on the ongoing generation of subspaces.

2.3 Words and Concepts in Context

The relationship between words and concepts is fraught, as Davidson (1974) points out: in the end, meaning, intention, and cognitive content seem to emerge in dynamic tension with one another, and it is therefore problematic to say that language in use can ever really stand for a deeper cognitive representation. On the other hand, language, and certainly natural language, can no more be the substrate of thought than it can sit on top of thought; it must be grasped as a thing in itself, pragmatic, often messy, and very much in the world. Putting words beside rather than above or below concepts, Clark (2006) interprets language as "a cognition-enhancing animal-built structure... a kind of self-constructed cognitive niche: a persisting but never stationary material scaffolding," (p. 370). So language can be seen as a thing in the world, a perceptible entity that, in the sense of affordance described by Gibson (1979), affords a linguistic agent the opportunity to do something conceptual. Words are not the same thing as the concepts which they serve to denote, nor are they the atomic substance of mental states which sustain concepts, but they are nonetheless instrumental in the unfolding phenomenon of conceptualisation in an environmental context.

Barsalou (1993) has defined "concepts" as "temporary constructions in working memory" (p. 34), and in particular he has examined what he identifies as the *haphazard content* of the inherently vague application of language to concepts. For Barsalou, concepts are specifically something other than *feature lists*, or mere enumerations of the things which perpetually constitute a concept. The prevalence of language in the representation of concepts assumes, though, that concepts are "stable structures in long term memory," (p. 44), and a consequence of is a fundamental *linguistic vagary* in the relationship between concepts and words. Barsalou considers the example of how words are used to conceptualise BIRDS: diverse features such as `feathers`, `nest`, and `eats worms` are drawn into the conceptual framework as demanded by a particular cognitive context. These momentary, contextual linguistic constructs can serve to delineate, among other things, cultural fault lines, with Chinese language speakers evidently considering `swan` and `peacock` as prototypes of BIRD and therefore associating the concept with a feature of `gracefulness` that is not salient in other cultural contexts. Barsalou describes this application of language to conceptualisation in terms of the construction of "*ad hoc* categories," (p. 32).

A similar vocabulary is employed by Carston (2010), who defines an *ad hoc concept* as one that "has to be inferentially derived on, and for, the particular occasion of use," (p. 158). Using language to communicate about concepts therefore involves a process of discovering without an excess of mental effort a cognitive scenario where the implications of an expression are satisfactorily coherent. And the key to conserving mental effort in the course of generating *ad hoc* concepts is to outsource cognitive work to the environment, which provides the context in which the mapping from language to concept works. Allott and Textor (2012) have a slightly different perspective, describing *ad hoc* concepts as being themselves individuated composites consisting of "activated information that is supposed to pertain to a category or property," (p. 201). Here, again, the critical element in the formation of concepts is that there is some sort of contextual process of engagement with a cognitive apparatus which results in the emergence of a new, situated information structure. For Carston, the concept is a consequence of a contextually sensitive implicature, where for Allott and

Textor the concept is a compound structure composed of information activated by a situation, but in both cases, the essential element in the formation of *ad hoc* concepts is contextual entanglement.

The work presented in this paper seeks to demonstrate a practical computational implementation of the kind of dynamic linguistic representations that allow for creative, contextual construction of conceptual spaces. The chief requirement for such an implementation is to build a language model which is likewise dynamic and contextually sensitive. So the key issue at stake here is the nature of the representations used in the language model: they must be interactive both with one another and with the input fed to the model. Thus we seek to build a model that captures what Barsalou (1993) has described as the "open-ended recursion and context dependence of linguistic representations" (p. 48) that facilitate the trafficking of *ad hoc* concepts. The best hope of doing this computationally would seem to be to associate the words which denote concepts and their features with mathematically tractable, geometrically expressible information structures, and so we next turn to a consideration of statistical and network based language models.

3. Distributional Semantic Language Models

Where Gärdenfors (2000) has described conceptual spaces in terms of latent dimensions that correlate to stimuli in the world, research in computational linguistics, including that of Widdows (2004), has developed a similarly spatial view of semantics geared more explicitly towards the domain of language. Models which represent word meanings as vectors in a vector space have been shown to have several advantages over more traditional, symbolic approaches in expressing continuous properties such as similarity and relevance, while still being able to be extended (at least in theory) to more complex semantic phenomena. For example, Widdows shows how a geometric view of meaning can be used to construct a multi-dimensional language model which, in turn, can be used to emulate higher level cognitive operations such as logic; Grefenstette and Sadrzadeh (2011) and Socher et al. (2012) offer alternative ways to compositionally construct sentence representations from their constituent words.

Clark (2015) offers a comprehensive overview of the field; in the following few pages, a brief history of recent developments in this area of computational linguistics will be offered, with a particular focus on the dimensional situation of these spaces.

3.1 Word Counting and Matrix Factorisation

One set of approaches relies on directly observed lexical statistics. In this view, usually termed *distributional semantics*, the fundamental premise is that similar words appear in similar contexts (Harris, 1957), and that semantics can therefore somehow be represented in terms of the way in which words are arranged in relation to one another across a corpus.

Motivated by statistical techniques for document retrieval such as Latent Semantic Analysis (LSA, Deerwester et al., 1990), Schütze (1992) proposed a method for representing the semantic relationships between words involving the construction of vectors based on co-occurrence counts observed across relatively large-scale textual corpora: Schütze's model built vectors populated by statistics indicating the frequency with which words occurred in the context of other terms, and then applied a clustering algorithm in order to locate angular regions in the space corresponding to senses of ambiguous words. Similarly, Lund and Burgess (1996) used a word-counting technique to build a model geared towards clustering words into conceptually oriented categories. Due primarily to concerns with the data storage and handling requirements entailed by very high dimensional

representations of words occurring in many different contexts, these authors employed dimensional reduction procedures to generate smaller, denser versions of their initial co-occurrence matrices. Schütze in particular used singular value decomposition, while Lund and Burgess picked dimensions with the highest variance, under the presumption that high variance corresponded to a high degree of informativeness across co-occurrences with a given context word. More recently, Rychlý and Kilgarriff (2007) use a range of heuristics to reduce processing complexity, but calculate full matrices to enable thesaurus creation.

Subsequent work involving statistical techniques for modeling meaning in terms of word co-occurrences across a large corpus has applied more complex mathematical approaches to representing the observed relationship between words. Blei, Ng, and Jordan (2003), for instance, presented a model which moves beyond mere word counting, seeking to model the "exchangeability" (p. 994) of words across different contexts in terms of probability distributions construed over a relatively small array of *latent* parameters. Expanding the scope of statistical language models in a different direction, Kanejiya, Kumar, and Prasad (2003) presented a model that applied singular value decompositions to matrices describing the co-occurrence of both the morphological forms of words and the syntactical dependencies between words, achieving marginal improvements over purely semantic models in correlation with human responses to a set of test questions. More recently, Turney and Pantel (2010) have suggested that, to the degree that a co-occurrence matrix might be construed as an incomplete register of the semantic possibilities inherent in a language, singular value decomposition can be viewed as "a way of simulating the missing text," (ibid, p. 160).

Notwithstanding the increasing sophistication of models, the theme which emerges across the relatively brief history of statistical text modeling (since at least LSA (Deerwester et al., 1990)) is one of dimensional reduction. The generation of dense and condensed matrices was originally driven by the necessity of computational efficiency, but latter day approaches have actually embraced dimensional reduction, and matrix factorisation in particular, as a mechanism for enhancing model performance. Pennington, Socher, and Manning (2014), for instance, describe matrix factorisation as an important element of their GloVe model, and Yogatama et al. (2015) report state-of-the-art results on word similarity, sentence completion, and sentiment analysis using a sparse coding technique for building nuanced distributional semantic models. Indeed, despite the mixed results of dimensionality reduction reported in the comprehensive survey of distributional semantic vector spaces by Lapesa and Evert (2013), an enthusiasm for matrix factorisation is currently prevalent in the field, with one recent pair of authors going so far as to assert that singular value decomposition type reduction "entails the abstraction of meaning by collapsing similar contexts and discounting noisy and irrelevant ones, hence transforming the real world term-context space into a word-latent-concept space which achieves a much deeper and concrete semantic representation of words," (Hassan and Mihalcea, 2011).

This notion of abstracting away from a dimension that literally corresponds to a co-occurrence observation is key to all the factorisation techniques described here. In this regard, the techniques predicated on the tabulating of word counting techniques that have just been described have a certain commonality with the arguably even more abstract *word embedding* approaches to vector space semantics, which will be described next.

3.2 Word Embeddings from Neural Networks

Around the same time as Blei, Ng, and Jordan (2003) were developing their nuanced statistical approach to topic modeling, Bengio et al. (2003) introduced an alternative methodology for modeling word meaning based on vector spaces built by multi-layer neural networks. The objective of this technique is similar to the statistically oriented work mentioned above: the construction of a space of word-vectors, where proximity corresponds to semantic information. The technique itself, however, is different, and this difference has been motivated by a stance on the very dimensionality of the space. To use the terminology of Bengio et al., the smooth redistribution of probability achieved by a neural network helps to overcome the *curse of dimensionality*, by which a small movement in a jagged space can lead to a catastrophic breakdown in the language model. This smooth distribution is achieved through the steadily annealed influence of the incrementally updated weights of a neural network, resulting in a space of word-vectors which have been described as *word embeddings*.

The early neural approaches to language modeling involved learning a function which mapped from input word-vectors to an output distribution assigning probabilities to the occurrence of a subsequent word, and in this sense were more in line with classic n-gram language models (Brown et al., 1992), though the underlying intuition was distributional, in that word-vectors for semantically similar words were expected to have similar features and therefore similar probabilities of occurring in a certain context. In subsequent work, Collobert and Weston (2008) have demonstrated a method for using a multi-layer network for assigning nuanced levels of features to words in a sentence, again relying on the premise that proximate vectors will be processed in similar ways by the network. Huang et al. (2012) developed a network which takes both local and non-local contextual information about a word as input, building a space of disambiguated representations where cosine distance between two sense specific word-vectors is explicitly taken as a measure of similarity.

Building on this body of work, Mikolov, Yih, and Zweig (2013) presented their `word2vec` model, based on a neural network that learned a space of word-vectors based on the distributional semantic insight, although rather than deriving vectors directly from co-occurrence statistics, the network training infers vectors which predict co-occurrence. The resulting model has geometric features especially suited for capturing analogical relationships between words. In this space, simple linear algebraic operations between word-vectors can yield meaningful results, with the paradigmatic example from the original literature being the calculus by which $\overrightarrow{woman} - \overrightarrow{man} + \overrightarrow{king} = \overrightarrow{queen}$. Two different and similarly effective approaches to training the vector space are presented, one involving a network that predicts a word based on its context and another that predicts a context based on a word. In both cases, the network weights and the values of word-vectors are simultaneously updated through backpropagation as the model uses a sliding context window to processes the corpus, much like with the word counting techniques described above. The result is a space loaded with semantic value: the actual geometry of the space yields significant information about the relationship between words, and this information might potentially be mapped onto conceptual schema.

With that said, all these neural language models essentially use dimensions as handles for gradually and systematically pulling word-vectors into a global arrangement which satisfies the semantic relationships observed between words, per the insight into contextual correlation of the distributional hypothesis. In fact, the random initialisation of matrices means that an entirely different space, albeit with similar relative relationships between word-vectors, will be established

for any given run of a model on a particular corpus. The dimensions themselves are thus populated by arbitrary values that cannot be interpreted as corresponding to any co-occurrence event in the underlying corpus. In this regard, the neural network models, like models based on matrix factorisation, seek to exchange expressiveness on a dimensional scale for compactness and robustness on the scale of the entire model. Despite this similarity, the recent trend in computational linguistics has been away from strictly statistical models and towards word embeddings, as characterised by the conclusions drawn by Baroni, Dinu, and Kruszewski (2014).

3.3 Finding Dynamic Context in a Lexical Space

The interesting – perhaps even remarkable – feature of the networks developed by Mikolov, Yih, and Zweig (2013) is that complex non-linear models underlie surface spaces where simple arithmetical operations between vectors reveal semantically loaded relationships between words, a point which has been explored by Arora et al. (2015), who go on to propose a generative model that, they suggest, restores an element of *interpretability* to their word-vectors. Levy, Goldberg, and Dagan (2015) have staged something of a counterattack on the ascendancy of word embedding approaches to language modeling, distilling what they deem to be *hyperparameters* deployed in the cases of the most effective systems, systemic features such as context windows of varying length that can, in principle, be applied to any model built through the traversal of large scale corpora. Once these modular model components are taken into account, as Levy et al would have it, the evident theoretical distinctions between the two general approaches to distributional semantics can be reduced to an array of testable technical considerations.

The upshot of this view is that there seem to be some grounds for considering both statistical and network based approaches to distributional semantics as building more or less the same kind of spaces through a variety of different techniques. Indeed, as discussed above, one of the main characteristics of the majority of distributional semantic models, regardless of how they are generated, is that their dimensions become abstractions that do nothing more than delineate a likewise abstract space. These are spaces just for the sake of being spaces, where any characteristic properties of the space itself are stripped out. In an early insight that is particular prescient with regard to the work presented here, however, Schütze (1992) observed that, with regard to distributional semantic models, "different dimensions are important for different semantic distinctions and that all are potentially useful," (p. 794). Schütze's point was that there is valuable information in the raw data of pointwise co-occurrence probabilities inherent in the comportment of words across a corpus, and this information could possibly become a strength of a sophisticated language model. This is a point that the model presented here aims to take seriously.

To return to Gärdenfors' insight regarding the geometric nature of conceptualisation, it seems impossible to imagine how a space defined by, typically, dozens to hundreds of purely abstract dimensions could ever be construed as containing the kind of conceptual differentiation that is evidently inherent in the relationship between minds and the world. This is not to say that a number representing the likelihood of two words occurring in the same context should be construed in the same richly interrelated and elaborately differentiated way as Gäredenfors' highly structured, complexly delineated conceptual spaces. Nonetheless, a space of numbers that serve both as anchors for the meaningful positioning of points and also as indicators that can be independently associated with events, even events which are themselves relatively abstract, is somehow more in the world than an essentially indivisible space of mere positions. In particular, in an unreduced base lexical space

of sparse, literal co-occurrence dimensions, there remains some hope of constructing contextualised projections by making an informed decision about which dimensions would best serve as the basis for these projections. The remainder of this paper will be devoted to describing and testing a model designed to do precisely that.

4. A Methodology for Conceptual Discovery

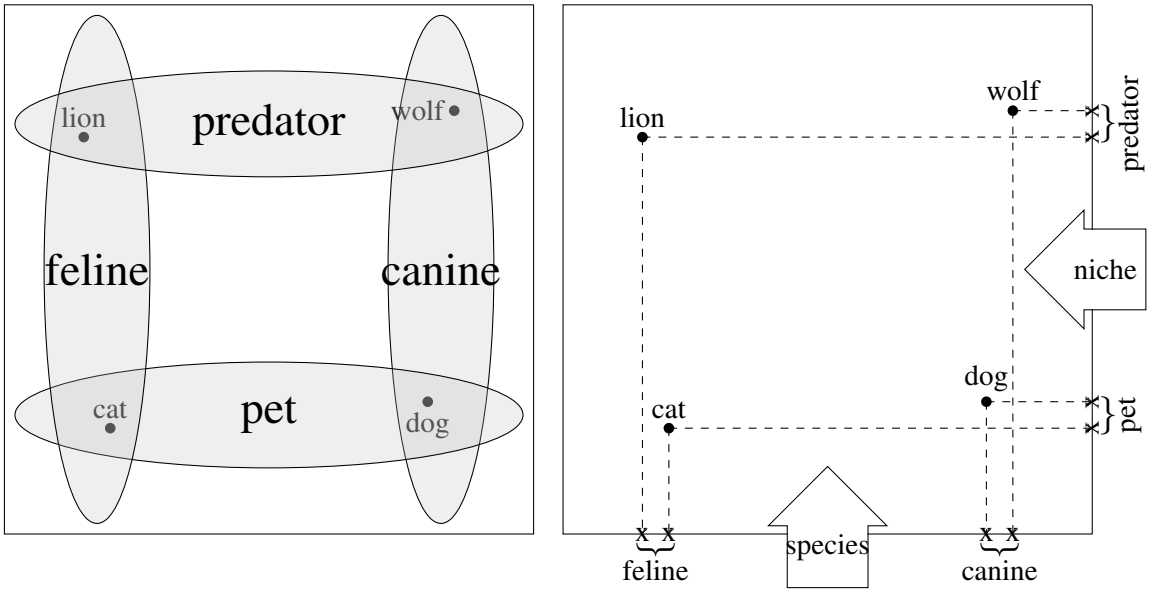
Our conceptual discovery model employs a contextually informed projection of a very high dimensional distributional semantic model into a subspace defined by features pertinent to the concept being queried. These characteristic features are derived from an analysis of a base set of word-vectors taken to be paradigmatic examples of the concept being sought. We predict that a subspace defined in terms of conceptually salient dimensions will remit a geometry of word-vectors that, when considered as a clustering of labeled points, can be mapped into a conceptual space.

4.1 A Distributional Semantic Space

In contrast to standard distributional semantics approaches, rather than simply dividing up a single lexical space in search of clusters that can be mapped to *ad hoc* concepts, we propose a technique for the *ad hoc* projection of a base space into a potentially vast array of context-specific, conceptually loaded subspaces. In general, a standard lexical space is laden with all the pragmatic messiness of language as used in functional communication. A word like `cat` is likely to be found in the proximity of terms associated with various contextual conceptualisations of CAT: with words like `dog` by virtue of its use to describe a HOUSEHOLD PET, with words like `lion` due to both its membership in and its use as a generic term for members of the FELINE conceptual category, and indeed with names like `Coltrane` because of its colloquial application as a term for constituents of JAZZ MUSICIAN. A minor example of this type of ambiguity is portrayed in Figure 1a. Such a general space is therefore not well suited to discovering conceptual spaces directly; it requires contextualisation in terms of a particular conceptual perspective.

In order to resolve the ambiguity and vagary inherent in lexical semantics, we propose performing dimensional reductions on the lexical space; however, rather than using a global technique like SVD as in standard NLP approaches, we select dimensions dynamically based on an analysis of the representational characteristics of the space's dimensions. The words in the lexical space are effectively projected into a lower order subspace based on a specific dimensional perspective, as portrayed in Figure 1b. We consider these perspectives on the lexical space to be analogous to projections into contextually informed conceptual spaces. In other words, we consider these reduced versions of the lexical space to map into a conceptual space, where clusters of words correspond to *ad hoc* conceptual regions. The base space we generate in the application of the method described in Section 5 is very high dimensional (in our model, it has approximately 7.5 million dimensions)—and it is precisely this vast dimensionality which provides a proliferation of different perspectives to be projected in subspaces. Every possible combination of dimensions corresponds to a different subspace, and therefore to a different potential mapping of word clusters into conceptual regions.

The existing models mentioned in Section 3 are designed to perform well on tasks involving compositional operations such as sentence construction and analogy completion, and yield state-of-the-art results for such tasks. The motivation behind these models is similar to our own: beyond simply exploring word similarity, they seek to discover relationships in the geometry of a semantic



(a) a lexical space

(b) a contextual projection

Figure 1: In the two-dimensional space depicted in (a), the conceptual vagary of four words maps to overlapping, elongated and indeterminate spaces. In (b), two different perspectives on the lexical space, represented by the arrows labeled *niche* and *species*, offer contextualised projections in one-dimensional clusters which remit conceptual clarity.

space. Our conceptual discovery model, however, attempts to do this in an *ad hoc*, query-specific way, generating subspaces based on contextual information inherent in a concept query. As this dynamic generation of query-specific spaces exploits the direct statistical relationship between initially unknown query terms and their co-occurrence with other terms in a corpus, our model must maintain a high-dimensional, markedly sparse space of literal co-occurrence statistics covering the entire vocabulary. (Baroni and Lenci, 2010, describe a model that uses a generative base matrix to generalise over a variety of task specific distributional semantic models.)

The model is based on a matrix M representing the co-occurrence of words with each other in context, as observed in some large corpus of language. M is a $p \times q$ matrix, where each row corresponds to a word w, and each column corresponds to a co-occurring term c. As such, the size of the model's vocabulary is p, and the number of co-occurrence terms is q. As in many standard distributional semantic models (see e.g. Clark, 2015), the matrix element $M_{w,c}$ is calculated as the positive pointwise mutual information between w and c:

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

where $n_{w,c}$ is the count of co-occurrences of word w with context term c, n_w is the count of occurrences of word w in any context, and n_c is the count of occurrences of context term c with any word, and W is the total count of all word tokens across the corpus.

The constant a is a smoothing term to mitigate the high mutual information value associated with very infrequent co-occurrence terms (the Laplace smoothing technique described by Turney

and Pantel, 2010). Adding 1 to the ratio of frequencies prior to calculating the logarithm guarantees that all values in the matrix will be positive: elements corresponding to w, c co-occurrences never observed will equal 0, and elements representing co-occurrences observed exactly as often as would be randomly expected will equal 1. As most words will never be observed co-occurring in most contexts, the matrix is largely sparse. And, as this is not a generative model, it is not appropriate to increment the zero-valued elements of the matrix in order to anticipate unobserved but still plausible co-occurrences (as common in language modelling: e.g. Manning and Schütze, 1999)—in fact, doing so would hamper the computation of the conceptually targeted subspaces, as explained below.

For any word w, we can now define the vector $\overrightarrow{w}$ that represents it in this space as the corresponding row of M, i.e. $\langle M_{w,1}, M_{w,2}, \ldots, M_{w,q} \rangle$.

4.2 Generating Ad Hoc Conceptually Targeted Subspaces

This very high-dimensional, largely sparse, distributional semantic space now serves as a base space for the generation of subspaces which cluster word-vectors in a way that is relevant to a particular concept, corresponding to the conceptual perspectives described in Section 4.1. These subspaces can be understood as a projection of the base space, calibrated for the geometric discovery of groups of words relevant to the perspective in question. The premise behind these projections is that an analysis of a small set of pertinent word-vectors should reveal a group of salient features that will indicate an effective set of dimensions for defining the desired subspace.

We assume as input a small set of n words relevant to a particular conceptual domain. This input might take the form of a set of descriptive terms (eg, `wild` and `animals`) or a set of categorical terms (eg, `lions`, `tigers`, and `bears`). We take the word-vectors for to each of the input terms, $\{\overrightarrow{w_1}, \overrightarrow{w_2} ... \overrightarrow{w_n}\}$ and extract from them the set of b features which have non-zero values across all n input word-vectors. We then project a new, dense $n \times b$ space, labeled N, of the n input word-vectors defined over the b non-zero features and normalise them, with each feature w_i^j of a vector $\overrightarrow{w_i}$ recalculated based on the norm of that vector. Although the standard Euclidean L2 norm can be used (equation 2), we have found that the L1 norm (equation 3) gives a slight advantage by sharing weight more evenly amongst dimensions (see below):

$$w_i^j = \frac{w_i^j}{\sqrt{\sum_{k=1}^{b} \left(w_i^k\right)^2}} \tag{2}$$

$$w_i^j = \frac{w_i^j}{\sum_{k=1}^{b} abs\left(w_i^k\right)} \tag{3}$$

In the base space M, vectors are not normalised, although the calculation of pointwise mutual information acts as a kind of normalising factor, with high frequency co-occurrences in the numerator of the informational equation mitigated by high independent observations for each word in the denominator. Once a subset of query vectors has been defined, however, normalisation allows us to smooth out any imbalance between terms, and this is more effective after selection of the non-zero features specific to this subset. For our n input words, once the features that are non-zero across all n word-vectors have been determined, it may turn out that some have significantly lower values for these non-zero features than the others. Normalisation ensures that all input word-vectors can contribute proportionally in a subsequent feature-by-feature analysis.

We now select a subset c' of the t most salient features for this set, determined as those with the highest mean value μ_c across all n word-vectors:

$$\mu_c = \frac{1}{n}\sum_{w=1}^{n} N_{w,c} \tag{4}$$

and with the salient features associated with the input word-vectors thus established, we use them to build the desired subspace by selecting only the features c'. This projects the original q-dimensional vectors in M into t-dimensional vectors, resulting in a new $n \times t$ matrix S, with rows defined by the input word vectors w and columns defined by the reduced set of conceptually salient features c':

$$M_{w,c} \Rightarrow S_{w,c'} \tag{5}$$

Our hypothesis is that the geometry of this projected space – defined *ad hoc* from the specific input terms – and in particular the relative spatial situation of word-vectors, should remit clusterings of words that can mapped, at least broadly, onto conceptual spaces. The final step in the process of discovering this conceptual word-space is to delineate the region within the space where these conceptually pertinent word-vectors are likely to be found.

4.3 Traversing the Subspace

We propose that, by projecting the dimensionally inclusive structure of M into a subspace S defined by dimensions strongly associated with the target concept, word-vectors corresponding to words with conceptually relevant characteristics will be drawn outward from the origin relative to other word-vectors in S. Because S has been defined in terms of features that characterise conceptually relevant input word-vectors – whether that input is descriptive or exemplary of the target concept – word-vectors corresponding to instances of the concept being sought should have high values for all the dimensions that delineate S. We therefore expect them to be situated fairly centrally in the positive region of the space (recalling that all values in this model are positive). With this in mind, we investigate two alternative methods for delineating and discovering concepts within S, which we term the *anchor* and *norm* methods.

Anchor method The first method defines concepts as regions around a point, which we term the *anchor point*. A natural candidate for this anchor point, which must be centrally located in S and have a relatively high norm, is a vector $\overrightarrow{v}$ located at the center of the positive region of the space:

$$\overrightarrow{v} = \langle \frac{1}{\sqrt{q}}, \frac{1}{\sqrt{q}}, \ldots, \frac{1}{\sqrt{q}} \rangle \tag{6}$$

In the present versions of the model, q is determined experimentally, but in general it should be a value on the order of the highest dimensional value found in Equation 5, as this will approximate the extent of S. We can now characterise any word-vector $\overrightarrow{w}$ in terms of its Euclidean distance d_w from the anchor point $\overrightarrow{v}$:

$$d_w = \sqrt{\sum_{c=1}^{q} \left(S_{w,c} - \frac{1}{\sqrt{q}} \right)^2} \tag{7}$$

and can derive a set of conceptually relevant word-vectors by simply taking the set R within some radius r of the anchor, i.e. with $d_w < r$. See Figure 2a for an illustrative example.

Norm method The second method relies purely on the fact that S is characterised by dimensions which are relevant to the query terms, and defines concepts as regions beyond a certain distance from the origin. In this approach, we therefore characterise a word-vector $\overrightarrow{w}$ in terms of its Euclidean distance d_O from the origin (L2 norm):

$$d_O = \sqrt{\sum_{c=1}^{t} (S_{w,c})^2} \tag{8}$$

See Figure 2b for an example. Intuitively, this is likely to discover more candidates for conceptual membership (tending towards higher recall), but might include more atypical or peripheral members (possibly tending towards lower precision).

4.4 Relation to conceptual spaces

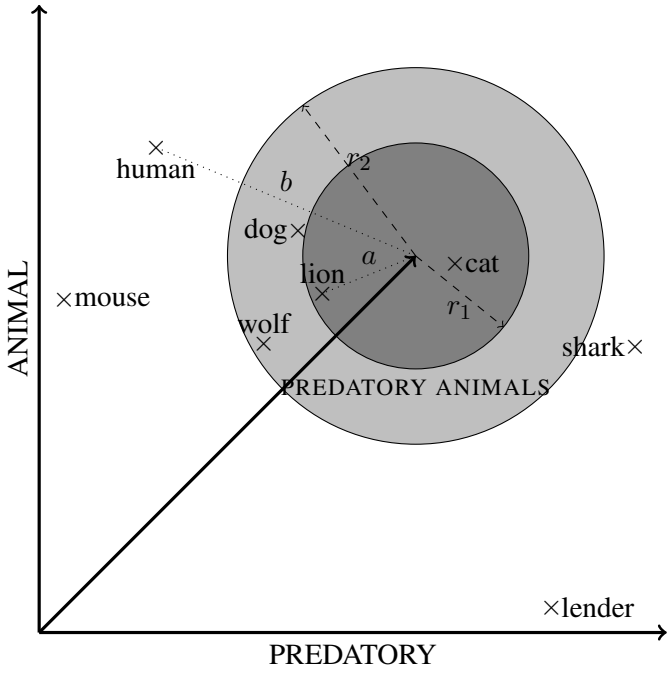
For both methods, we take the word-vectors in the defined conceptual region as candidate instances of the concept being queried: they will be the terms most strongly associated with the selected dimensions, which are in turn the co-occurrence features most strongly associated with the joint input terms. The provision of at least two input terms provides a crucial degree of contextualisation; otherwise we would effectively be performing a search for terms that occur in similar contexts to a single word-vector, reducing our model to a standard dimensionally reduced approach to word similarity tasks.

As well as the ability to incorporate specific, contextual information, we see our approach as providing a lexical model which is commensurate with the main properties of Gärdenfors (2000)'s conceptual spaces. The subspace S is relatively low-dimensional, compared to the original standard co-occurrence space. It is also characterised by dimensions which are interpretable in terms of observed co-occurrence in language, as opposed to standard dimensionally-reduced distributional models (e.g. Milajevs et al., 2014), and low-dimensional neural embeddings (e.g. Mikolov, Yih, and Zweig, 2013), which provide low dimensionality but thus lose interpretability. Finally, the conceptual region is convex: in the anchor method, it is a standard Euclidean hypersphere; in the norm method, the region is essentially characterised uni-dimensionally, with betweenness defined in terms of norm, guaranteeing convexity.

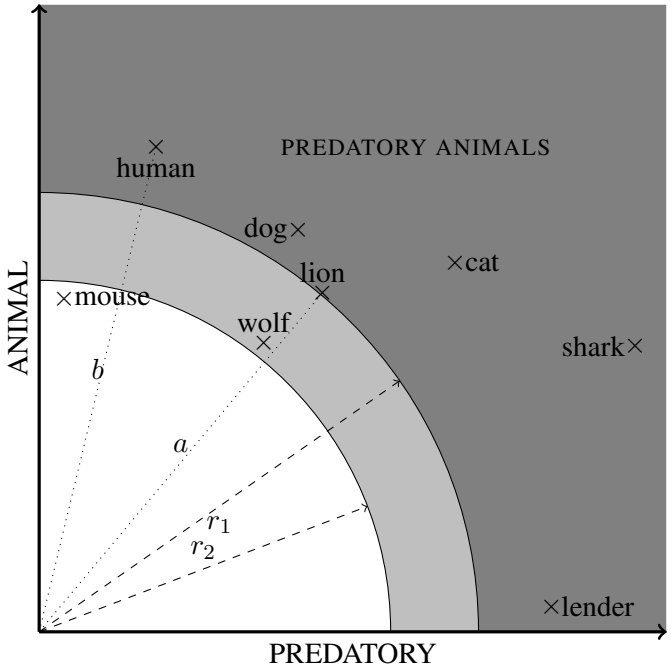
5. Study 1: Concept Discovery in WordNet

Our first evaluation of this approach is on a task of conceptual membership discovery, using WordNet (Fellbaum, 1998) as our ground truth, Wikipedia as our co-occurrence corpus, and comparing against two leading standard distributional semantic models: one taking the neural-net approach, `word2vec` (Mikolov, Yih, and Zweig, 2013), and one based directly on co-occurrence statistics, GloVe (Pennington, Socher, and Manning, 2014) (see Sections 3.1 and 3.2). Our hypothesis is that, given in particular compound input queries denoting relatively high level WordNet synsets, our model should perform at least as well at recapitulating the set of terms that are members of the corresponding WordNet subtree as `word2vec` and GloVe.

In evaluating by comparing against an existing ontology, we follow existing work on conceptual modeling through distributional semantics such as that presented by Cimiano, Staab, and Tane (2003), who develop a formalism for defining concepts in terms of logical sets with shared members



(a) Anchor method. The concept PREDATORY ANIMALS is delineated by a circle centered on the unit vector of equal positive elements: the circle with radius r_1 captures words associated with paradigmatic instances of the concept, while the circle described by r_2 begins to pick up less obvious constituents. Distances such as a and b are the actual metrics for determining conceptual membership.



(b) Norm method. The concept PREDATORY ANIMALS is delineated by an arc centered on the origin: the region with radius $r > r_1$ captures words associated with paradigmatic instances of the concept, while the more central region $r > r_2$ begins to pick up less obvious constituents. Distances such as a and b are the actual metrics for determining conceptual membership.

Figure 2: A sampling of data from the space described in Section 5 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.

and attributes, populating these using syntactically informed dependency information about the co-occurrence of verb-object pairs observed in a domain-specific corpus. They compare the output against the structure of an existing ontology, using metrics of *lexical overlap* and *semantic cotopy* to measure the accuracy of their model both syntactically and semantically. Likewise, Snow, Jurafsky, and Ng (2006) combine a range of techniques to induce a taxonomy, evaluating against WordNet; their approach has a similar theoretical objective to our own in its attempt to overcome the ambiguity inherent in many conceptual labels. Finally, Baroni, Dinu, and Kruszewski (2014)'s comprehensive survey of statistical and word embedding methodologies for building distributional semantic language models includes a set of *conceptual categorisation* tasks involving clustering a predefined vocabulary into groups associated with likewise predefined conceptual categories; evaluation involved measuring the percentage of correct classifications made for each category.

In contrast to these tasks, which included learning categories for simple terms, our motivation is in conceptualisation for novel, compound terms and conceptual blends, to explore our model's ability to project contextually informed subspaces. As such, we use a similar test based on recapitulation of the members of a set of WordNet hypernyms, but with a specific evaluation set: a set of high level nodes in WordNet's hierarchy that are denoted by compound labels, as described below.

5.1 Dataset and Training

For these experiments, we trained our model and the two established models on the English language Wikipedia.[1] We pre-processed the text by disregarding non-sentential elements such as section headings, labels, captions, lists and references; removing punctuation; rendering all text into lowercase characters; and only considering sentences of at least five words in length. This procedure yielded a corpus of approximately 1.1 billion word tokens distributed across just under 7.5 million word types. In all cases, we built models based on a vocabulary of the 200,000 most frequently occurring word types in the text, with the exception of the articles `the`, `a`, and `an`, meaning that all word types in the vocabulary appeared as tokens at least 83 times in the corpus. We experimented with with symmetrical context windows of both 4 words (2 on either side of a target word) and 10 words (5 on each side of the target word) – see (e.g. Lund and Burgess, 1996; Pennington, Socher, and Manning, 2014) who use similar values.

Constructing the Distributed Semantic Space For our model, training consists of deriving the basic co-occurrence matrix M from observations of words within the given context window. We recorded a total of slightly over 1 billion occurrences of the top 200,000 tokens (corresponding to N in Equation 1—the top 0.025% of types account for the vast majority of tokens in a distribution with a very long tail), and a total of about 9.4 billion tokens of the 7.5 million context word types (keeping in mind that the same contextual token might co-occur in several different contexts in the same sentence). We calculated the mutual information for every co-occurrence based on Equation 1, determining that a value of 10,000 was appropriate for the smoothing constant a via trial and error.

We trained `word2vec` using the Gensim package for Python.[2] We experimented with both the *CBOW* technique, which learns vector representations for predicting words based on contexts,

1. The December 8, 2014 dump, downloaded from http://meta.wikimedia.org/wiki/Data_dump_torrents#enwiki on January 23, 2015, parsed into plain text using the "Wikipedia Extractor" software, downloaded from http://medialab.di.unipi.it/wiki/Wikipedia_Extractor on February 13, 2015.
2. Downloaded from https://radimrehurek.com/gensim/.

and the slightly less conventional *Skip-gram* technique, which learns to predict contexts based on an input words. We tested across 300 dimensional spaces built from 1, 3, and 10 iterations over the corpus using symmetrical context windows of 4 and 10, always setting negative sampling to 10, the initial learning rate to 0.025, and downsampling to 1×10^{-5}. These values were derived from previous experiments discussed by Mikolov et al. (2013) and Baroni, Dinu, and Kruszewski (2014).

For the GloVe model, we used the glove-python package developed by GitHub user maciejkula.[3] We trained 100, 200, and 300 dimensional spaces on 10, 20, and 50 iterations across co-occurrence matrices based on symmetrical context windows of 4 and 10 words. These values were primarily derived from the primary literature (Pennington, Socher, and Manning, 2014), though through preliminary testing we found that, despite reports of marginal improvement between 50 and 100 dimensional models, for our tests the lower dimensional models actually worked better. The explanation for this may be that the model is overfitting at higher iterations on the relatively small corpus we used.

5.2 Procedure

Our task here is to discover members of concepts, given the names of those concepts as seed terms. We take WordNet *synsets* to correspond to concepts, with the words they contain in the form of lemmas corresponding to the members of those concepts. As our motivation here is towards the discovery of novel concepts, rather than standard dictionary concepts, we take as our evaluation set the subset of WordNet synsets which have compound, two-word names. We consider the top 40 such synsets with the highest number of global children nodes in the WordNet hierarchy, and then, within this subset, only those synsets which are not parents of other synsets in the selected group of highly populated synsets, resulting in a total of 23 target conceptual categories. Our logic here is that this process will select highly populated synsets but will also discriminate against those very high level synsets which contain overly general concepts such as LIVING THING. The resulting queries, along with results for top performing models, are tabulated in Appendix A.

To compare the models, we use the two words from the synset name as input, and evaluate precision as the proportion of the top 50 output terms of each model which are lemmas of the members of the target synset on WordNet. In all cases, we only consider model output which corresponds to a unique morpheme as returned by `nltk's`[4] implementation of a WordNet stemmer. This gives a measure of precision; given the nature of the WordNet synsets here, which contain large numbers of words including very rare words, recall seems a less useful measure for this task. For our proposed model, the top output terms are selected on the basis of distance either from the origin (the *norm* model) or the central point (the *anchor* model) of the subspace S. For `word2vec` and GloVe, the top output terms are selected on the basis of mean cosine distance from the two input terms, the standard method from Mikolov, Yih, and Zweig (2013) for discovering relevant similar terms.

All models, including `word2vec` and GloVe, are trained on the Wikipedia dataset discussed above. We test each candidate model (our normed model, our anchor model, word2vec, and GloVe) across a range of parameters (including context window size and number of dimensions – see above). We then select the best performing parameter settings for each model, so that no model is advantaged or disadvantaged by choosing a particular set (this approach for comparing

3. Downloaded from https://github.com/maciejkula/glove-python.
4. Python's Natural Language Toolkit package, found at http://www.nltk.org/.

computational models has precedence; see for example Pearce, Müllensiefen, and Wiggins, 2010). For our norm model, this selection yields a model based on 300 dimensions and a context window of 5 terms. Our anchor model also uses a context of 5, but with 200 dimensions. For word2vec, the best performing model uses CBOW, trained over a single iteration using a context window of 2 terms; the best GloVe model likewise builds a 300 dimensional model using a context window of 2, trained over 10 iterations.

Example In order to illustrate our experimental procedure, we will consider a single case of how each model handles an input query. We take as our instance the concept BODY PART, a high level synset node in the WordNet hierarchy. We take as our target set all lemmas of all members of the subtree with the BODY PART node as its root—which is to say, all words considered to be hyponyms of `body part` by WordNet. We then supply `body` and `part` as input to our model and both of the established models. In the case of our model, we perform the dimensional analysis described in Section 4.2 on the word-vectors $\overrightarrow{body}$ and $\overrightarrow{part}$ as found in our base lexical space. We then search the subspace projected from this analysis for both those word-vectors with the highest norm and those vectors closest to the central anchor point described in Section 4.3, with q in Equation 6 set to 5, a value we determined through trial and error. In the case of `word2vec` and GloVe, we calculate the mean vector between the two input word-vectors and then search for the word-vectors with the lowest cosine differences from this mean vector in the respective spaces. In all cases, we compare the top 50 results with unique morphemes in WordNet's vocabulary and compare them with the target set, calculating a precision statistic accordingly.

5.3 Results and Discussion

A comparison of performance across models highlights several interesting findings regarding the *types* of concepts that lead to differential model performance on the conceptual membership task, as well as the slipperiness of delineating certain types of concepts (in both linguistic and computational settings). Taking a holistic view of the results, we see that all models perform modestly for a number of input queries – our norm model achieves precision of greater than 15% (more than 7 hits out of 50 terms) for half of the queries (12 of 23), and word2vec achieves greater than 15% precision for only 10 of 23 queries. Precision quantifies how well the models generate appropriate terms (hits) given all of the output generated, formally defined as the fraction of output terms, out of the 50 generated terms, that are found in that particular WordNet subtree.

On average, both our concept discovery model (with the norm method) and `word2vec` outperform GloVe, as shown in Appendix A. Our norm model appears to give the highest average precision, and a t-test shows no statistically significant difference between the performance of our model and `word2vec` ($p > 0.8$ in two-sided and matched-pairs t-tests), demonstrating that our model performs as well as `word2vec` on this task for the given set of concepts. Comparing our models' performance, the norm method outperforms the anchor method. Because the performance of GloVe and our anchor model is worse than our norm model and `word2vec`, our subsequent analyses will focus on the latter two models. Interestingly, however, our model performs better at certain kinds of concepts and worse at others in comparison to `word2vec`.

To examine differential performance across concepts, we divided the 23 queries from WordNet into three categories of concepts prior to data analysis. The first category includes very specific and/or scientific concepts, including DICOT GENUS, TELEOST FISH, REPRODUCTIVE STRUCTURE, and BIRD GENUS. The spaces projected from an analysis of the dimensional overlap between these

pairs might predictably be dominated by one especially specific element of the query. In the case of input terms like `dicot` and `teleost` in particular, which probably occur infrequently in our base corpus and in very particular contexts, we expect to find very sparse vectors with co-occurrence data concentrated in a small number of dimensions. It is also worth noting that with some of these queries, the lopsidedness of the projection produced corresponds to an informational unevenness in the query itself: `fish` adds very little to `teleost`, since the latter is more or less explicitly a specification of the former.

The second class of concepts were those which are very broad or vague, including GROUP ACTION, KNOWLEDGE DOMAIN, FUNDAMENTAL QUANTITY, SOCIAL GROUP, PHYSICAL PHENOMENON, ORGANIC PROCESS, and CHANGE OF STATE. These queries are characterised by pairs of terms in which both constituents are general and would be expected to have a broad, flat co-occurrence distribution, corresponding to relatively dense populations of low but non-zero values in their word-vectors. The spaces projected should therefore be correspondingly general, and the model as it is currently constructed might struggle to project an adequately contextualised space.

The third classification type were those occupying the conceptual middle ground – concepts which are contextualized and somewhat specific in terms of meaning, but also not too narrow in scope. This list included ILL HEALTH, BODY PART, WOODY PLANT, WRITTEN COMMUNICATION, ORGANIC COMPOUND, TIME PERIOD, CONSUMER GOODS, NATURAL LANGUAGE, AUDITORY COMMUNICATION, SKILLED WORKER, MONETARY UNIT, and TRANSFERRED PROPERTY. In these instances, we find terms such as `body` and `part` which might individually be quite broadly distributed in the corpus, but the intersection of their co-occurrences as discovered in the course of dimensions should be relatively informative. In terms of the relationship between words and concepts, this corresponds to a successful, highly informative contextualisation.

When examining the performance of both our model and word2vec on the above concepts, it is clear that performance varies across these three categories of concepts, with both models displaying differential performance according to the type of concept. This finding is more pronounced for our model, with Specific concepts garnering a precision rate (across the constituent concepts) of 0.030, Broad concepts displaying precision of 0.062, and Moderate concepts yielding superior performance with a precision of 0.306. In comparison, word2vec precision results are more moderately consistent across the three categories, with precision rates at 0.146 for Specific concepts, 0.082 for Broad concepts, and 0.242 for Moderate concepts.

Both models demonstrate superior performance on Moderate categories that are neither broadly or vaguely construed, nor specific or technical. As predicted, our model does especially well for concepts which are contextualized, in the sense that both terms contribute to the meaning of the combined query. This aspect of our model might be interpreted in terms of informational entropy: while our word-vectors are not, strictly speaking, probability distributions, they are based on probabilistic information derived from observations of word co-occurrence. From this perspective, it seems that our model does particularly well when both terms in a query pair have similar, and perhaps similarly broad, distributional characteristics. To put this in information theoretical terms, concepts labeled with word-vectors with relatively low entropy should be particularly well. Returning to the theoretical analysis offered above, such queries correlate with concepts that are

This analysis might also be construed in terms of abstraction, and in this regard our model relates to work from Hill, Korhonen, and Bentz (2014) assessing the influence of word associations on human reports of abstractness versus concreteness of concepts. Those authors found in particular that concepts considered to be more abstract tended to have a broader, flatter distribution of

associations to other concepts. They furthermore draw a connection between abstraction and *association*, as opposed to *similarity*, with the premise being that groups of concrete concepts tend to be clustered in terms of similarity (so, DOG is similar to CAT), while abstract concepts are grouped in terms of association (DOG is associated with LEASH). Returning to the distributional hypothesis, we might then hypothesise that a distributional semantic language model would place concrete words in proximity to one another, since semantic similarity is predicted to correspond to the contextual similarity that these models exploit. With a similar interpretation in mind, Hill et al propose that distributional models that use syntactic information as an element of their feature dimensions (such as the one described by Kanejiya, Kumar, and Prasad (2003) and Padó and Lapata (2007)) should be well equipped to capture more abstract relationships between concepts.

It is interesting, then, to note that our model seems to thrive in a middle ground between abstraction and concretion, despite there being no syntactic information built into our word-vectors. This seems like a good indicator of the way our selective dimensional reduction technique finds a subspace where words can be clustered in a conceptually productive way: by being particularly selective about the context we use to define each projection, our methodology draws some of the associative information inherent in abstraction into the model. The key observation to made here for the time being is that, based on our model's output, there is good reason to look for a relationship between conceptual abstraction and the dimensionality of the word-vectors that correspond to the concept's label.

On the above test of conceptual membership, our model performs better than (in the case of GloVe), or is statistically indistinguishable from (in the case of Word2Vec), state-of-the-art methods in distributional semantics, justifying our model's place among the cutting-edge methods used in the field. This finding is striking, because the comparison models are more complex than our own, and our model is able to capture conceptual spaces in which the dimensions and geometry of the spaces themselves are meaningful. The query terms afforded from WordNet are limiting, however, because our model has been constructed for discovering conceptual spaces reflecting *ad hoc* concepts. The structure of the WordNet ontology does not allow for the comparison of a wide range of contextualized or uncommon concepts (for example, "Poetic Creativity"), which is where we predict our model to do quite well. This limitation of input queries motivates further exploration of contextualized concepts, using queries not found in WordNet. Because this leaves us without a ground truth against which to interpret model performance, we use empirical testing, by having participants list members of provided contextualized concepts.

6. Study 2: Empirical validation of concepts

Study 1 shows that our approach performs at least as well as standard distributional models on average, and is better suited to some concept types which are our primary interest, given our overall motivation in concept modelling and computational creativity. The conceptual queries in Study 1 were heavily constrained, however, by the structure and idiosyncrasies of WordNet (there were only 23 concepts in total that met our criteria for testing). In addition, several of those were not truly contextualized or blended concepts, because one term effectively encompassed the other. Consider, for example, TELEOST FISH – there are no examples of teleosts that are not a fish, and therefore, the second term is effectively redundant and does not serve to contextualize or constrain the concept. Given that our interest here is in the ability to delineate *ad hoc* concepts, the set of queries afforded by WordNet is insufficient for our purposes. To this end, we conducted further testing on conceptual

terms that are not present in WordNet, but that are true combinations of terms, not simply cases in which one word carries most of the semantic information for the pair. To provide a ground truth for these, human participants were asked to generate candidate terms during a concept association task, and these were compared with the output of the vector space models obtained as before.

Study 2 therefore sought to test a range of conceptual categories spanning a subjective range of concreteness/abstraction, such as WILD ANIMAL and INFECTIOUS DISEASE. We do not quantify the level of abstraction of queried terms (as our aim is not to delve into the philosophical debate concerning conceptual abstraction); rather, the selected concepts were chosen to provide a more varied test of model performance across different types of concepts. Human participants were asked to generate three terms that they associate with each concept provided. In this study, a total of eight concepts were presented during the online conceptual membership task, and the terms provided by participants were analyzed for comparison with the top output terms from our distributional model using both the anchor and norm techniques, as well as from the top performing version of `word2vec`.

In Study 1 we evaluated using a measure of precision: of the models' top output terms, the proportion included in the relevant WordNet subtree (with the large number of candidates in that subtree making a recall measure inappropriate). In this study, we focus on *recall*: of the human responses for a concept, the proportion matching outputs from the model. Precision would be a less useful metric here, as the number of human responses is limited; reasonable model outputs might therefore be missing from the human list. In this study, we used the standard L2 norm to normalise our conceptual subspaces, as per Equation (2); differences in performance using the L1 norm are small. In the case of both the norm and anchor methods, we projected 200 dimensional spaces based on a context window of 5 words.

6.1 Method

6.1.1 PARTICIPANTS

Thirty-five participants (avg age = 40.1 yrs, stdev = 14.3 yrs) volunteered to take part in the study, of whom 24 were female.

6.1.2 PROCEDURE

An online Qualtrics questionnaire was used to collect responses from volunteers. After giving informed consent, participants were asked to list three terms they associate with a particular concept. The queried concepts included WILD ANIMAL, BRIGHT COLOR, CULINARY CREATIVITY, INFECTIOUS DISEASE, POSITIVE EMOTION, POLITICAL IDEOLOGIES, PROFESSIONAL OCCUPATIONS, and THEATRICAL CREATIVITY. These eight concepts were presented in randomized order for each participant. After submitting their responses, participants provided basic demographics information about their age, ethnicity, native language, and gender. Lastly, participants were debriefed on the goals of the study.

The human responses were tabulated and then stemmed using `nltk's` Porter Stemmer, resulting in a list of unique stems. Each model's output was likewise process, resulting in a list of 50 independent base terms that could be compared against the human considerations of each conceptual query.

theatrical & creativity	political & ideologies	bright & color	wild & animal	infectious & disease	professional & occupations
artistry	liberalism	greenish	deer	encephalitis	technologist
musicality	ideology	pinkish	goats	meningitis	careers
musicianship	marxists	brownish	hares	autoimmune	licensure
showmanship	capitalism	bluish	foxes	measles	dentists
ingenuity	populism	grayish	civets	syphilis	tradesmen
choreographic	nationalism	tint	raccoons	dengue	pharmacists
spontaneity	espousing	tinge	bison	hemorrhagic	self-employed
sophistication	anti-capitalist	hues	antelopes	rabies	apprenticeships
inventiveness	ideals	blotches	ungulates	herpes	nurses
puppetry	syndicalism	orange-red	skunks	respiratory	electricians
improvisational	marxist-leninist	blue-green	hyenas	pathogenesis	credential
artists	authoritarianism	iridescent	boars	sepsis	dietitians
stagecraft	radicalism	underparts	rhinoceros	gastrointeritis	physiotherapists
individuality	secularism	colouration	nilgai	leprosy	plumbers
evocative	extremism	greyish	blackbuck	leishmaniasis	part-time

Table 1: The top 15 output vectors for six conceptual queries from Study 2, as returned by our anchor point technique.

6.2 Results and Discussion

This study compares model output with human terms associated with the eight queried concepts. In terms of overall recall, the norm version of our model scores **0.16** across the 8 concepts, the anchor version scores **0.23**, and for comparison, word2vec scores **0.14**. Therefore, as mentioned above, we focus on the anchor version of our model for the discussion of results below. Results are listed category-by-category in Table 2.

As an illustration, the top 15 output terms (out of 50 used in the study) using the *anchor* model are listed in Table 1, for six of the 8 concepts tested in Study 2. For comparison with the output of our model, participants' responses were merged into an exhaustive list for each of the eight queries. These terms were then compared with the top 50 output vectors of each conceptual space discovered by the model. The terms in common between participants and the anchor model included the following:

WILD ANIMAL: `bear`, `wolf`, `meat`, `feral`, `lion`, `tiger`, `foxes`, `endangered`, `habitat`, `boar`, `zoo`

BRIGHT COLOR: `yellow`, `orange`, `blue`, `green`, `red`, `pink`, `white`, `light`, `paint`, `vivid`, `pure`

CULINARY CREATIVITY: `innovation`, `taste`, `unique`, `culture`, `cooking`, `combining`, `experience`, `beauty`, `imagination`, `sense`

INFECTIOUS DISEASE: `hepatitis`, `measles`, `influenza`, `syphilis`, `virus`, `illness`, `typhoid`, `contagious`, `vaccine`, `outbreak`, `communicable`, `smallpox`

POSITIVE EMOTION: `empathy`, `feeling`, `optimism`, `satisfaction`

	norm	anchor	w2v
WILD ANIMAL	0.10	0.20	0.14
BRIGHT COLOR	0.04	0.22	0.10
CULINARY CREATIVITY	0.23	0.17	0.14
INFECTIOUS DISEASE	0.17	0.20	0.09
POSITIVE EMOTION	0.09	0.11	0.13
POLITICAL IDEOLOGIES	0.21	0.23	0.16
PROFESSIONAL OCCUPATIONS	0.19	0.35	0.11
THEATRICAL CREATIVITY	0.16	0.29	0.22
overall	0.16	0.23	0.14

Table 2: Recall statistics measuring the output of two versions of our model and one version of `word2vec` against human output for eight conceptual categories. Recall is calculated by dividing the number of matches with the human output by the total number of human outputs. The overall recall is the ratio of total correct hits in all categories to the total number of human responses.

POLITICAL IDEOLOGIES: `socialism`, `liberalism`, `anarchism`, `communism`, `leftist`, `fascism`, `libertarian`, `extremists`, `capitalism`, `totalitarianism`, `conservatism`, `idealism`, `democracy`

PROFESSIONAL OCCUPATIONS: `profession`, `accountant`, `nurse`, `education`, `consultant`, `administration`, `vocation`, `salary`, `careers`, `lawyer`

THEATRICAL CREATIVITY: `artistic`, `musicals`, `visual`, `innovative`, `dancing`, `production`, `performance`, `inspiration`, `expressive`, `costumes`, `originality`, `brilliant`, `stage`, `improvisation`, `artistic`, `show`, `performance`, `beauty`, `emotion`, `inspiration`, `entertainment`

Overall, a compelling set of results emerge. In the case of our anchor point technique in particular, all categories were recalled at a rate of at least 10%, with considerably higher rates for a number of queries, in particular PROFESSIONAL OCCUPATIONS and THEATRICAL CREATIVITY. It is also interesting to note differences in performance between each space defining technique: the norm method did not do very well at finding exemplars of BRIGHT COLOR, meaning that the projection entailed by the dimensional analysis of `bright` and `color` failed to find a space where terms that humans thought were relevant were marked by the highest overall dimensional values. More generally, it is worth noting that, where the norm method proved superior in the precision task described in Study 1, the anchor method has done better in recapitulating this smaller set of human responses. This might be explained in terms of the prototypicality of the responses we received from our participants: where the categorisations of WordNet aim for some degree of comprehensiveness, the sampling of responses we have here may pertain more to paradigmatic views of conceptual constituency, and this in turn is better modeled by searching a central region of a contextually projected lexical subspace.

As with Study 1, the performance of `word2vec's` word embeddings seems to be more even, but also, in the case, notably worse than that of our model. It would seem that the character of the queries used for this second study are more in line with the contextually interpretable types of conceptualisations that our model has been designed to handle, where two terms with comparable

degrees of ambiguity and informativeness serve to contextualise one another in a denotation of a concept that is not overt from either term independently. Per the analysis offered in Section 5.3, we postulate that the concepts featured in Study 2 better capture the level of generality and abstraction that our model is equipped to process.

7. Study 3: Direct Comparison of Model and Human Terms

The model appears capable of discovering highly relevant output for the above concepts, but simply conducting a comparison of human- and model-generated terms does not permit an assessment of the quality of terms discovered by the model which humans did *not* cite. Also, it is likely that if more human responses were collected, more responses would be in common with the model's output. To address this issue of insufficient comparison, and to provide a more direct analysis of model output, a third study was conducted in which the top 20 output vectors from each query in Study 2 were evaluated alongside human-generated terms.

In both the case of model output and of participants' responses, some terms may be considered more exemplary of the concept while others are more peripherally related. Therefore, in addition to explicitly evaluating model output, this study was also intended to elucidate whether *human*-generated terms are generally seen as belonging to the concept they were intended to illustrate. This will provide a baseline for comparison with the results for our model.

7.1 Method

7.1.1 PARTICIPANTS

Seven participants (avg age = 36.3 yrs, stdev = 13.1 yrs), including 1 female and six males, volunteered to take part in the study.

7.1.2 PROCEDURE

For each of the eight conceptual queries from Study 2, participants were given a list of 40 terms to rate, including the top 20 output vectors from the model interspersed with 20 terms provided by humans. The 20 human-generated terms were randomly selected from the exhaustive list of terms collected for each concept. The participants' task was to mark with an "X" those terms that they believe belong to the indicated concept (which was noted above the list of 40 terms). All terms were listed in randomized order for each of the eight concepts in question, and participants were not informed of the source of generation, as this can bias evaluation of human output compared with computationally-generated output (Moffat and Kelly, 2006).

7.2 Results and Discussion

We sought to test whether the model's output is comparable to human-generated terms for each query. If this is the case, there should be no statistical difference between the number of human- and model-generated terms rated as belonging to a given concept. Across all eight conceptual queries, the overall percentage of human terms rated with concept membership was 61.8%, while the overall percentage for model-generated terms was 53.3%. The model's terms were rated comparably to human terms, but the number of model terms judged as representative of a concept depended on the type of concept being tested. To investigate these results, a two-tailed t-test was performed for

each of the 8 queried concepts, to assess whether there was a significant difference in the raters' responses between the human- and computer-generated terms. The dependent variable for each test was the proportion of individuals rating each term as belonging to the given concept. Because we are interested in whether there is a difference in how participants rate human versus model terms, we test whether the null hypothesis (i.e., that there is no difference between the two sets of terms) holds for each query. If the distributions for the two sources of data are not statistically different (and therefore the null hypothesis is not rejected), this gives support for a claim that the model's output is not rated differently from human output. A full claim of equivalence would require a different test; future work may construct tests for specific comparisons.

Of the eight concepts tested, six tests failed to reject the null hypothesis, supporting the claim that for these concepts, human- and model-generated terms were not rated differently. These six concepts were WILD ANIMAL (t = -1.95, p = 0.06), POLITICAL IDEOLOGIES (t = -0.80, p = .43), CULINARY CREATIVITY (t = 1.70, p = .10), PROFESSIONAL OCCUPATIONS (t = -1.96, p = 0.06), INFECTIOUS DISEASE (t = 0.8, p =.39), and THEATRICAL CREATIVITY (t = -0.27, p = 0.79). There were also two cases in which the null hypothesis was rejected, namely, POSITIVE EMOTION ($t = 5.96, p < 0.01$) and BRIGHT COLOR ($t = 5.04, p < 0.01$). For these two queries, human terms were more often judged to be members of the associated concept than model terms.

A few interesting conclusions may be drawn from these findings. First, the model performs reasonably well for a range of concepts, as human and computer terms produced similar ratings for concrete concepts such as WILD ANIMAL and more abstract concepts such as POLITICAL IDEOLOGIES and THEATRICAL CREATIVITY. Also, the two queries for which the model did not perform comparably to humans, POSITIVE EMOTION and BRIGHT COLOR, are arguably the two concepts tested that are used in the widest range of linguistic contexts. In the case of POSITIVE EMOTION, there is such widespread use of emotion-terms in language that our method of using pointwise mutual information produced terms which are less frequently used in many contexts, such as the slightly more psychological examples of "appraisals" and "cognitions". In the case of BRIGHT COLOR, there is a similar emphasis on biologically-relevant terms, such as "colouration" and "underparts". Additionally, although any color may be considered bright or dark, there is likely an enculturation effect at play, such that the concept BRIGHT COLOR more reliably evokes exemplars such as orange and yellow than gray or brown. Our results speak to how functional use of some color terms may differ quite substantially from others.

8. Summary and Future Directions

In this paper, we use a multidisciplinary approach, employing methods from computational linguistics as well as empirical psychology, to describe a system for discovering concepts and their members. The model takes arbitrary conceptual terms as input to map into corresponding conceptual spaces: starting with a very high-dimensional space based on lexical co-occurrence, we use the input terms to find vastly reduced subspaces corresponding to these *ad hoc* concepts, and discover both the dimensions which characterise them, and the conceptual sets which they delineate.

We evaluated this model in several ways: In Study 1, we performed a rigorous comparison between two versions of our model and two of the leading models in distributional semantics, `word2vec` and GloVe. After a comprehensive parameter search, we discuss the best models of each type, based on their performance on a conceptual membership list completion task. For the 23 input queries derived from WordNet, we found both our model and `word2vec` perform better on

average than GloVe, while no statistical difference was found between the norm measuring version of our model and `word2vec`. This places our system among the highest-performing models in the field for this task, with the advantages of being less complex than the comparison models, while discovering spaces whose geometry is more compatible with Gärdenfors' approach to conceptual spaces in its view of concepts as convex regions within subspaces with semantically meaningful dimensions.

Because the concepts in the first study were heavily constrained by the idiosyncrasies and synset structure of WordNet, and as our model was predicted to do well on contextualised concepts, further testing of our model was warranted with a more broad range of concepts. To this end, Study 2 explored a range of contextualised concepts, which varied subjectively in terms of abstraction and linguistic functionality, to test the robustness of our distributional semantics vector space model. Our model outperforms word2vec, with the conceptual regions discovered producing semantically relevant terms, many corresponding directly to participants' terms, and others extending the list of terms to insightful new dimensions.

Study 3 confirmed that the model can discover semantic categories and indices of concepts that are aligned to human conceptualizations. In all but two of the concepts explored, there were no significant differences between human ratings of computational output and human terms. This said, the model did not capture all of the semantic categories cited by humans. For example, in its definition of creative domains, our model evinced a notable omission of language pertaining to emotion, particularly of concern as terms relating to affect and evoked emotional response were some of the most frequently cited for these concepts. Accordingly, future work will investigate why the model does not capture this aspect of the conceptualisations in question.

Further directions for the future include the application of this computational approach to contextually-defined conceptual domains, both for the ontologically useful task of elaborating concepts themselves, and to create well-tailored terminology for the assessment of creative output from the corresponding domains. This methodology may also be used to more directly approach the task of conceptual blending: rather than specifying input vectors that belong to only one concept, one may supply input dimensions from several. This could result in output terms discovered at the intersection of the lexical regions specified by the vectors' different input dimensions.

Another direction for future work is exploring ways in which adjusting various parameters of the system might reveal a more fine-grained conceptual delineation of language. For instance, refining the technique for exploring contextually projected subspaces could reveal regions of different relationships, where some clusters of words might be considered examples of a concept while other clusters might map to properties of a concept. A practical application of such features of the space could be, for instance, the automatic generation of taxonomies. Other parameters to be explored include the dimensionality of the subspaces, the metrics for picking these dimensions, the context window size for building the base lexical space, and the corpus used for generating the space.

Ultimately, we would like to apply the model to the production of more explicitly creative artefacts. Here, again, the strength of our model should be its inherently geometric character: by mapping concepts as regions in a lexical subspace, the chance for connections between conceptual domains based on congruence between word clusters arises, and this in turn should provide a platform for modeling the formation of conceptual metaphor. In terms of the theoretical applications of our model, we feel that the approach we've outlined here goes some way towards offering a practical implementation of a philosophical stance on the nature of conceptualisation, and in this regard, the model is intended not only to perform well on quantitative computational linguistic

metrics, but also to provide a noteworthy method for building representations that are fungible and dynamic. This, we maintain, is the essence of creativity: an ability to incorporate the ongoing emergence of unpredictable context into a flexible conceptual framework which results in the construction, through the composition of representations, in interesting and useful new conceptualisations.

Acknowledgments

The first author is supported by EPSRC grant EP/L50483X/1. The remaining authors' contribution is funded by the Lrn2Cre8 and ConCreTe Projects, which acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grants number 610859 and 611733, respectively.

Appendix A. Study 1 data table

model	NORM		ANCHOR		CBOW		SKIP-GRAM		GLOVE	
window	5	5	5	5	2	5	5	5	2	5
dimension	300	200	200	100	300	300	300	300	300	100
iteration	-	-	-	-	1	1	1	3	10	20
WOODY PLANT	10	4	3	3	2	3	2	3	4	4
SOCIAL GROUP	2	3	12	10	11	13	6	17	4	10
BODY PART	39	41	15	12	18	16	14	6	13	12
ILL HEALTH	30	29	15	13	14	11	9	9	7	5
DICOT GENUS	1	1	0	0	7	6	8	1	27	18
WRITTEN COMMUNICATION	17	19	22	21	12	14	15	14	13	11
ORGANIC COMPOUND	11	12	0	1	19	17	21	18	2	3
GROUP ACTION	2	2	5	6	7	7	9	10	12	12
KNOWLEDGE DOMAIN	9	10	5	6	5	4	3	4	2	5
FUNDAMENTAL QUANTITY	0	0	5	5	2	2	1	3	3	3
TIME PERIOD	8	4	15	17	28	29	15	17	5	17
CONSUMER GOODS	9	10	4	6	8	7	5	6	8	3
NATURAL LANGUAGE	20	21	4	2	21	16	6	6	9	0
AUDITORY COMMUNICATION	1	0	2	1	1	1	0	0	0	2
TELEOST FISH	3	2	1	2	19	20	22	15	13	13
PHYSICAL PHENOMENON	5	4	3	4	2	4	2	1	1	2
SKILLED WORKER	22	21	4	4	18	12	14	16	14	10
ORGANIC PROCESS	0	0	1	1	0	0	0	0	1	0
REPRODUCTIVE STRUCTURE	2	2	0	0	2	3	2	2	0	0
CHANGE OF STATE	4	4	4	2	2	5	4	5	0	1
BIRD GENUS	0	0	0	0	1	2	1	2	1	0
MONETARY UNIT	9	14	1	1	1	1	0	0	1	1
TRANSFERRED PROPERTY	8	7	7	8	3	4	3	4	2	3
precision	0.184	0.183	0.111	0.109	0.177	0.171	0.141	0.138	0.123	0.117

Table 3: Category by category precision results for Study 1, comparing model output to all lemmas associated with members of assorted WordNet subtrees labeled with compound terms. The top two results for each model are presented: our model's output measured by greatest norm and proximity to a central anchor point, plus `word2vec's` CBOW and Skip-Gram methods and GloVe. The numbers associated with each category are the number of hits scored by each model's top 50 output terms, and the final column is each model's precision over all 23 conceptual categories.

References

Allott, N., and Textor, M. 2012. Lexical Pragmatic Adjustment and the Nature of Ad Hoc Concepts. *International Review of Pragmatics* 4(2).

Arora, S.; Li, Y.; Liang, Y.; Ma, T.; and Risteski, A. 2015. Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings. *CoRR* abs/1502.03520.

Baroni, M., and Lenci, A. 2010. Distributional Memory: A General Framework for Corpus Based Semantics. *Computational Linguistics* 36(4):673–721.

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 238–247. Baltimore, Maryland: Association for Computational Linguistics.

Barsalou, L. W. 1993. Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of Compositional System of Perceptual Symbols. In Collins, A. F.; Gathercole, S. E.; Conway, M. A.; and Morris, P. E., eds., *Theories of Memory*. Hove: Lawrence Erlbaum Associates. 29–101.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Jauvin, C. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3:1137–1155.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Boden, M. A. 1990. *The Creative Mind: Myths and Mechanisms*. London: Weidenfeld and Nicolson.

Brown, P. F.; deSouza, P. V.; Mercer, R. L.; Della Pietra, V. J.; and Lai, J. C. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4):467–479.

Carston, R. 2010. Lexical Pragmatics, Ad Hoc Concepts and Metaphor. *Italian Journal of Linguistics* 22(1):153–180.

Cimiano, P.; Staab, S.; and Tane, J. 2003. Automatic acquisition of taxonomies from text: FCA meets NLP. In *In Proceedings of ECML/PKDD Workshop on Adaptive Text Extraction and Mining*.

Clark, A. 2006. Language, Embodiment, and the Cognitive Niche. *Trends in Cognitive Sciences* 10(8).

Clark, S. 2015. Vector Space Models of Lexical Meaning. In Lappin, S., and Fox, C., eds., *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell.

Collobert, R., and Weston, J. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25 th International Conference on Machine Learning*.

Davidson, D. 1974. On the Very Idea of a Conceptual Scheme. In *Proceedings and Addresses of the American Philosophical Association*, volume 47, 5–20.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Jounal of the American Society for Information Science* 41(6):391–407.

Fauconnier, G., and Turner, M. 1998. Conceptual Integration Networks. *Cognitive Science* 22(4):133–187.

Fellbaum, C., ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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

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

Grefenstette, E., and Sadrzadeh, M. 2011. Experimental Support for a Categorical Compositional Distributional Model of Meaning. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Harris, Z. 1957. Co-Occurrence and Transformation in Linguistic Structure. *Language* 33(3):283–340.

Hassan, S., and Mihalcea, R. 2011. Semantic Relatedness Using Salient Semantic Analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Hill, F.; Korhonen, A.; and Bentz, C. 2014. A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science* 38:162–177.

Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume Long Papers: Volume 1, 873–882.

Kanejiya, D.; Kumar, A.; and Prasad, S. 2003. Automatic Evaluation of Students Answers using Syntactically Enhanced LSA. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, 53–60.

Koestler, A. 1964. *The Act of Creation*. London, UK: Hutchinson.

Lapesa, G., and Evert, S. 2013. Evaluating Neighbor Rank and Distance Measures as Predictors of Semantic Priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Levy, O.; Goldberg, Y.; and Dagan, I. 2015. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3.

Lund, K., and Burgess, K. 1996. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. *Behavior Research Methods, Instruments, and Computers* 28(2):203–208.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICLR Workshop*.

Mikolov, T.; Yih, W.-T.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 246–251.

Milajevs, D.; Kartsaklis, D.; Sadrzadeh, M.; and Purver, M. 2014. Evaluating Neural Word Representations in Tensor-Based Compositional Settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 708–719. Doha, Qatar: Association for Computational Linguistics.

Moffat, D., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. In *Proceedings of the International Joint Workshop on Computational Creativity*.

Padó, S., and Lapata, M. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics* 33(2):161–199.

Pearce, M. T.; Müllensiefen, D.; and Wiggins, G. A. 2010. The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception* 39(10):1367–1391.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Rychlý, P., and Kilgarriff, A. 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 41–44. Prague, Czech Republic: Association for Computational Linguistics.

Schütze, H. 1992. Dimensions of Meaning. In *Proc. ACM/IEEE Conference*, 787–796.

Snow, R.; Jurafsky, D.; and Ng, A. Y. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 801–808. Sydney, Australia: Association for Computational Linguistics.

Socher, R.; Huval, B.; Manning, C. D.; and Ng, A. Y. 2012. Semantic Compositionality Through Recursive Matrix-vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, 1201–1211. Stroudsburg, PA, USA: Association for Computational Linguistics.

Turney, P. D., and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* (37):141–188.

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

Wiggins, G. A. 2006. A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. *Journal of Knowledge Based Systems* 19(7):449–458.

Yogatama, D.; Faruqui, M.; Dyer, C.; and Smith, N. A. 2015. Learning Word Representations with Hierarchical Sparse Coding. In *Proceedings of the 32nd International Conference on Machine Learning*.