Stage Two Report:
A Geometric Method for Context Sensitive Distributional Semantics

by

Stephen McGregor

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

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

July 2017

Abstract

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

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

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

computational linguistic practice.

Glossary

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

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

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

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

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

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

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

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

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

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

Table of Contents

Chapter 1

Conceptual Clusterings, Similarity, and Relatedness

In Chapter **??**, I laid out the theoretical groundwork for statistical context sensitive models of lexical semantics, and in Chapter **??** I described the actual methodology for building such much. In this chapter, I will now present the first set of experiments designed to evaluate the utility of this methodology. These experiments are intended to probe the productivity of a context sensitive, geometric approach to building a computational model of semantics based on statistics about word co-occurrences. They encompass two different experimental set-ups and corresponding varieties of data, one of which has been designed specifically for the purpose of testing my ideas and one of which involves an assortment of data used pervasively by computational linguistics interested in semantic models.

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

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

1.1 A Proof of Concept

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

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

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

			JOINT		INDY		ZIPPED	
	SG	BoW	norm	dist	norm	dist	norm	dist
top-50	6.230	6.584	8.726	6.208	7.725	5.014	9.311	5.774
full	3.413	3.634	4.144	3.203	3.686	3.005	3.945	3.093

and so *leg* would be considered a positive output for the input *body part*.[1]

1.1.1 Experiment Set-Up

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

All lemmas associated with all hyponyms of each synset are extracted and grouped. The terms labelling a given synset are then passed to my model, with the corresponding word-vectors serving as the basis for dimensional selection using the JOINT, INDY, and ZIPPED techniques as outlined in Chapter **??**. The subspaces returned by each of these techniques are explored to return the top terms using both of the procedures outlined in Chapter **??**: the terms closes to the mean point between the input word-vectors in a subspace are returned, and the terms furthest from the origin – the terms with the largest norm – in a given subspace are returned. The top 50 terms found in a subspace each according to each measure are returned, as well as the top n

1.1.2 Results and Analysis

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

References

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

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

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

Barsalou, L. W. (1993). *Theories of Memory*, Chapter Flexibility, Structure, and Linguistic Vagary in Concepts: Manifestations of a Compositional System of Perceptual Symbols. Hove: Lawrence Erlbaum Associates.

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* London: Jason Aronson Inc.

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin (2003). A neural probabilistic language model. *Journal of Machine Learning Research 3*, 1137–1155.

Birkhoff, G. (1958, 05). Von neumann and lattice theory. *Bulletin of the American Mathematical Society 64*, 50–56.

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

Bullinaria, J. A. and J. P. Levy (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods 44*(3), 890–907.

Burgess, C. and K. Lund (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive processes 12*(2/3), 177–210.

Carnap, R. (1947). *Meaning and Necessity: A Study in Semantics and Modal Logic.* University of Chicago Press.

Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. In *Proceedings of the Aristotelian Society*, Volume CX, pp. 297–323.

Casasanto, D. and G. Lupyan (2015). All concepts are ad hoc concepts. In E. Margolis and S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts.* Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origins, and Use.* New York, NY: Praeger.

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

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

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990).

Indexing by latent semantic analysis. *Journal for the American Society for Information Science 41*(6), 391–407.

Derrac, J. and S. Schockaert (2015). Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning. *Artificial Intelligence 228*, 66–94.

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines 22*(2), 87–99.

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

Erk, K. and S. Padó (2008). A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 897–906.

Erk, K. and S. Padó (2010). Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 Conference Short Papers*, pp. 92–97.

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

Evans, V. (2009). *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press.

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

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

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press.

Geffet, M. and I. Dagan (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 107–114.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and Semantics Volume 3: Speech Acts*, pp. 41–58. New York: Academic Press.

Gutiérrez, E. D., E. Shutova, T. Marghetis, and B. K. Bergen (2016). Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

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

Jäger, G. (2010). Natural color categories are convex sets. In M. Aloni, H. Bastiaanse, T. de Jager, and K. Schulz (Eds.), *Logic, Language and Meaning: 17th Amsterdam Colloquium, Amsterdam, The Netherlands, December 16-18, 2009, Revised Selected Papers*, pp. 11–20.

Kalchbrenner, N., E. Grefenstette, and P. Blunsom (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Kaplan, D. (1979). On the logic of demonstratives. *Journal of Philosophical Logic 8*(1), 81–98.

Kartsaklis, D. and M. Sadrzadeh (2016). Distributional inclusion hypothesis for tensor-based composition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December*

11-16, 2016, Osaka, Japan, pp. 2849–2860.

Kay, P. and L. Maffi (1999). Color appearances and the emergence and evolution of basic color lexicons. *American Anthropologist 101*(4), 743–760.

Kiela, D. and S. Clark (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) @ EACL 2014*, Gothenburg, pp. 21–30.

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

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

Landauer, T., D. Laham, B. Rehder, and M. E. Schreiner (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 412–417.

Lapesa, G. and S. Evert (2013, August). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, Sofia, Bulgaria, pp. 66–74. Association for Computational Linguistics.

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

Levinson, S. C. (2001). Yélî dnye and the theory of basic color terms. *Journal of Linguistic Anthropology 10*(1), 3–55.

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

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

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

Mikolov, T., W. tau Yih, and G. Zweig (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*, pp. 246–251.

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

Montague, R. (1974). English as a formal language. In R. H. Thompson (Ed.), *Formal Philosophy: selected papers of Richard Montague.* New Haven, CT: Yale University Press.

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

Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce.* Harvard University Press. edited by Charles Hartshorne and Paul Weiss.

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

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

Rimell, L. (2014). Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg.

Rorty, R. (1979). *Philosophy and the Mirror of Nature.* Princeton University Press.

Russell, B. (1905). On denoting. *Mind 14*(56), 479–493.

Sahlgren, M. *Italian Journal of Linguistics 20.*

Salton, G., A. Wong, and C. S. Yang (1975). A vector space model for automatic indexing. In *Proceedings of the 12th ACM SIGIR Conference*, pp. 137–150.

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

von Neumann, J. (1945). First draft of a report on the edvac. Technical report.

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

Widdows, D. (2003). Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pp. 136–143.

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

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence 11*(3), 197–223.

Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered Sets*, Dordrecht/Boston, pp. 445–470. Reidel.

Wille, R. (2005). *Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies*, pp. 1–33.