What is the Dimensionality of Human Semantic Space?

Will Lowe

Center for Cognitive Studies
Tufts University
Medford, MA 02155, U.S.A.
wlowe02@tufts.edu

Abstract.

McDonald and Lowe [15] showed that cosines in a semantic space of several hundred dimensions reflect human priming results for a wide range of semantic and associatively related words [16, Exp.2]. Previously, Lowe [11, 10] argued that the intrinsic dimensionality of semantic space is much lower, and that high-dimensional structure can be effectively captured in just two dimensions as the surface of a neural map. This paper provides a replication of McDonald and Lowe's results in two dimensions using the Generative Topographic Mapping [2], a statistically motivated neural network architecture for topographic maps.

1. Semantic Space

Semantic space model have proved very successful models of semantic memory [9, 3]. A semantic space operationalizes the idea, initially introduced in distributional linguistics, that words are semantically similar to the extent that they behave in the same way in text; verbs that subcategorize for the same sorts of arguments and nouns that can be modified by the same kind of adjectives are to that extent semantically similar. Semantic space representations use vectors of surrounding word counts as a substitute for knowing the distributional profile and argument structures in advance [12]. Words with similar vector representations share more similar linguistic contexts and are thus more semantically similar.

The success of semantic space models in modeling psycholinguistic phenomena to a large extent vindicates the approach to meaning underlying distributional linguistics, but it also raises a number of technical questions relating to the 'non-parametric' nature of the approximations made when constructing a space. If argument structures are known in advance then it is obvious how to measure similarity: for any two words look in each argument slot and compare the sorts of words found there. Thinking of all words as having subcategorization preferences may not be immediately intuitive but is explicit in Link and Dependency grammar [8], and helps connect semantic space work to syntactic perspectives. When argument frames are not known we count all surrounding words up to a maximum window size. The question is then: how many words do we need to count to get a good approximation? Semantic spaces are vector spaces of typically high dimensionality, so the the generalization of this question addressed in this paper is: what is the appropriate dimensionality of human semantic space?

2. Previous Work

Landauer and Dumais [9] have argued that there is an optimal number of dimensions for psychological modeling, and that data of appropriate dimension should be generated by taking large numbers of word counts and subjecting them to linear dimensionality reduction. Their claim is that human semantic space is of fairly low dimension compared to the dimensionality of vector data from a semantic space model. They estimate [9, Fig.3] that the 300 directions of principal variance should be retained from the thousands generated by their model to optimally predict human behaviour.

The idea that the intrinsic dimensionality of data is typically lower than its observed dimensionality motivates Multidimensional Scaling (MDS) and Factor Analytic approaches in psychology. MDS performs a similar function to neural models of topographic map formation in computational neuroscience [5, 7]. With the development of the Generative Topographic Mapping (GTM; [2]), a non-linear extension of Factor Analysis, these models are now not usefully distinguished. Ritter and Kohonen [17] used a self-organizing map to project simple vector representations of word co-occurrence counts onto a two dimensional map surface. Similar approaches have been taken by Scholtes [18] and Lowe [11]. Implicit in this work is the assumption that co-occurrence data is inherently very low-dimensional.

3. Interpreting Semantic Space Models

There are two distinct ways to interpret semantic space models. A space may be a description of the lexical semantic structure of a language. In this sense, constructing a semantic space is a methodology for finding semantic structure in English using a distributional similarity measure. Alternatively a semantic space may be a theory of semantic representation in people. On the first interpretation when distances in a space correlate reliably with human performance on some psychologically interesting measure we can infer that there is sufficient statistical regularity in the linguistic environment to be able to perform the psychological task. However, for a computational approach to psychology this is only half the story; there needs to be another theory of how that information is represented in the mind/brain. Semantic spaces *can* be psychological models: e.g. we might assert that each person has vectors of lexical associations and performs similarity computations on them to determine semantic similarity. However, this interpretation is not the one being tested when semantic distances are correlated with a human experimental performance. When the Hyperspace Analogue to Language (HAL; [14]) or Latent Semantic Analysis [9] is compared to human data there is no analysis by subjects, only by items. This is true of most previous work in semantic space. There are no subjects; we are testing a theory about items.

The work reported below treats neural network models as subjects and thus doubles as a theory of semantic representation, as well as a theory of the intrinsic dimensionality of semantic space itself.

The next section briefly reviews earlier work modeling associative and multiple types of semantic priming using a semantic space of high-dimension. The next section shows how substantially the same results are obtained if the dimensionality of the data

is reduced dramatically. Finally we consider the implications of this work for estimating the dimensionality of human semantic space.

3.1 Experiment 1: Priming in High-dimensional Space

Moss and colleagues [16] showed that semantic priming occurs for a wide range of semantic relations, both with and without association. Stimulus words named members of the same taxonomic category (category coordinates), either natural objects or artifacts, or they were related functionally (functional items), through script or instrument relations. Moss and colleagues showed separate semantic and associative priming effects for all categories. They also showed that the semantic priming effect was greater in the presence of association (the associative boost).

McDonald and Lowe [15] demonstrated that Moss *et al.*'s results can be modeled in a high dimensional space. We briefly review the details of model construction and results using the latest version of the model for comparison to the low-dimensional results described below.

We constructed a semantic space from 100 million words of the British National Corpus (BNC), a balanced corpus of British English [4]. Word vectors were generated by passing a moving window through the corpus and collecting co-occurrence frequencies for 536 of the most reliable context words within a 10 word window either side of each stimulus item. Context words were the same as those used in previous work modeling graded and mediated priming [13]. The method for choosing reliable context words is described elsewhere [12, 13]. We used positive log odds-ratios to measure the amount of lexical association between each context word and each of the experimental stimuli. The odds ratio is well-known to be a measure of association that takes chance co-occurrence into account [1].

We also created vectors for 1000 filler words of frequency ranks 1000 to 2000 in the BNC (114.55 to 49.15 occurrences per million). Stimulus frequencies ranged from 0.02 to 1639.23 per million, with a median frequency of 33.95 per million. 114 unrelated primes were chosen randomly from the set of filler words

As in the original experiment we varied three factors: association (associated, non-associated), semantic type (category coordinate, functional relation) and relatedness (related, unrelated). Semantic subtypes were nested under Semantic Type.

For the purposes of modeling priming, the cosine between a prime and target should be inversely proportional to the corresponding reaction time. The size of a priming effect is calculated by subtracting the cosine between the unrelated prime and target from the cosine between the related prime and target. Cosines for the unrelated prime-target pairs was taken to be the cosine of the target with another prime in the same condition. Cosines are entered directly into analyses of variance.

3.2 Results

Cosines in the semantic space are shown in Table 1. There was a main effect of relatedness, $\mathrm{F}(1, 108) = 314.922$, $p < .001$, and of association, $\mathrm{F}(1, 108) = 16.433$, $p < .001$, replicating associative priming. There was also an interaction between association and relatedness, $\mathrm{F}(1, 108) = 9.939$, $p < .01$. This replicates the associative boost.

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.553	0.173	**0.379**	0.458	0.163	**0.295**
Functional	0.547	0.181	**0.366**	0.394	0.168	**0.226**

Table 1: Cosines from the high-dimensional semantic space with unrelated primes chosen randomly from an alternative source. Bold face numbers are priming effects for each semantic type.

We then considered the associated and non-associated items. Semantic priming occurred in the associated condition, $\mathrm{F}(1, 54) = 205.972, p < .001$ and in the non-associated condition, $\mathrm{F}(1, 54) = 113.309, p < .001$. The priming effect for category coordinates appeared slightly larger than for functional items which is also consistent with the human results, but this difference was not significant.

Among the category coordinates semantically related pairs were more similar than unrelated pairs, $\mathrm{F}(1, 52) = 165.567, p < 0.001$. There was also associative priming, $\mathrm{F}(1, 52) = 5.607, p < 0.05$. There was no associative boost, $\mathrm{F}(1, 52) = 2.623, p = .111$, and no other significant interactions. The associative boost did not occur due to a low level of similarity between the associated artifact targets and their related primes. The delicacy of the boost also follows human results.

Functional pairs also showed a semantic priming effect, $\mathrm{F}(1, 52) = 154.771, p < .001$, a main effect of association, $\mathrm{F}(1, 52) = 11.555, p < .01$, and a reliable associative boost, $\mathrm{F}(1, 52) = 8.661, p < .01$. There was also a main effect of subtype, $\mathrm{F}(1, 52) = 4.58, p < .05$, due to steadily decreasing amounts of similarity across subtypes relative to a stable baseline (associated related script $>$ associated related instrument $>$ non-associated related script $>$ non-associated related instrument).

Detailed analyses for each semantic subtype are reported elsewhere [12].

3.3 Discussion

The space replicates Moss *et al.*'s finding that semantic priming occurs for a wide range of semantic categories, with and without association. We also see an associative boost.

3.4 Experiment 2: Priming in Low-dimensional Space

In this experiment we use 20 GTM networks as subjects. 20 GTM models [2] were trained on 1689 transformed semantic space vectors. The large number of irrelevant word vectors were intended give each network a better idea of the overall shape of semantic space, rather than just a small set of words of interest. 1000 words were the

filler from Experiment 1, 224 were from Moss's materials, and the rest were experimental stimuli from 5 other priming experiments. Results from the latter are reported elsewhere [13, 12].

The entire augmented set of 534-dimensional semantic space vectors was then transformed linearly into 50 dimensions by 20 independently generated stochastic matrices [6], one for each GTM model. Each GTM was initialized with random parameters and saw a distinct random mapping of the semantic space vectors.

Ideally each network would have been trained on vectors generated by sampling from a much larger corpus. However, this is computationally extremely demanding, even were such a corpus available. Using newsgroups is a possible next step in this research.

3.5 Random Mapping

Neural networks are often criticized for relying crucially on intelligent prior tranformations of the data. Consequently although principal component analysis of the vectors would also reduce dimensionality to tractable levels, it would also represent a substantial modeling assumption that is not obviously motivated from a neural perspective. Random mapping reduces the dimensionality of the data to a level that is tractable for reasonable network training times while making the fewest possible assumptions about the nature of the input, save that it derives from vectors of lexical associations.

Random mapping also introduces variability into the input data that ensures that no net trains on the same data set. The psychological interpretation of this process is that networks are subjects that have been exposed to roughly the same language data but with significant amounts of noise. We then test the claim that representing this information using topographic maps generates accurate predictions about priming.

Any specific random mapping for the semantic space vectors into d-dimensional space is a 534 $\times$ d matrix, $\mathbf{R}$, with i.i.d. zero mean Normally distributed elements. Each column of $\mathbf{R}$ is normalized to unit length to create a non-orthogonal basis. To give an idea of how much structure is preserved in a random mapping, Kaski [6] has shown that the inner product between two low dimensional projections $\mathbf{a}_1 = \mathbf{R}\mathbf{h}_1$ and $\mathbf{a}_2 = \mathbf{R}\mathbf{h}_2$ of high-dimensional unit length vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ is

$$\mathbf{a}_1^\mathsf{T}\mathbf{a}_2 = \mathbf{h}_1^\mathsf{T}\mathbf{h}_2 + \delta \tag{1}$$

where δ is approximately $\mathcal{N}(0, 2/d)$. This result essentially defines an error bar on similarity estimates in the low-dimensional space relative to their 'real' values in high-dimensions[1]. It is intuitively surprising that similarities are (on average) so well preserved by a completely random mapping; this phenomena alone deserves further attention.

3.6 Generative Topographic Mapping

The GTM is a non-linear extension of Factor Analysis with strong similarities to the Self-organizing map. It attempts to build a generative model of the variance structure in

[1] See Kaski's paper for error estimates for non-normalized high-dimensional vectors, and a detailed derivation.

data points $\mathbf{a}$ on the assumption that they are generated by a smooth non-linear mapping from a two-dimensional manifold $\mathbf{x}$. The GTM thus explicitly assumes that the intrinsic dimensionality of the data is two-dimensional, and that off-manifold structure is simply noise. Clearly this is an extremely strong and falsifiable assumption to make about even 50-dimensional data. Since the GTM defines a mapping from a low-dimensional latent space into the data space, it is straightforward to invert this mapping for any data point to obtain $p(\mathbf{x} \mid \mathbf{a})$. The mean of this distribution is a point estimate of the point in latent space that is most likely to have generated $\mathbf{a}$. In this respect the model is used similarly to Factor Analysis.

To make specific predictions about priming effects, we compute posterior means as described above for each related prime, unrelated prime and target vector. Each mean is a two element vector that describes a point in two-dimensional space. We take cosine measures in this reduced space, just as in the high dimensional model.

3.7 Results

	Associated			Non-associated		
	Related	Unrelated	U-R	Related	Unrelated	U-R
Cat. Coord.	0.689	-0.151	**0.839**	0.485	-0.250	**0.735**
Functional	0.690	-0.113	**0.803**	0.440	-0.134	**0.574**

Table 2: Mean cosine similarity measures from the networks on Moss *et al.*'s data with independently chosen unrelated baseline. Bold numbers are priming effects for each semantic category, with and without association.

Mean similarity measures are shown in Table 2. There was a main effect of relatedness, $\mathrm{F}_1(1,19) = 2391.276$, $p < .001$, $\mathrm{F}_2(1,108) = 195.478$, $p < .001$. There was also a reliable effect of association, $\mathrm{F}_1(1,19) = 94.703$, $p < .001$, $\mathrm{F}_2(1,108) = 7.703$, $p < .01$, replicating the associative priming effect. The associative boost was significant by subjects, $\mathrm{F}_1(1,19) = 21.316$, $p < .001$, $\mathrm{F}_2(1,108) = 1.949$, $p = .166$.

There were main effects of semantic relatedness in both associated and non-associated conditions, $\mathrm{F}_1(1,19) = 2358.761$, $p < .001$, $\mathrm{F}_2(1,54) = 103.68$, $p < .001$ and $\mathrm{F}_1(1,19) = 643.53$, $p < .001$, $\mathrm{F}_2(1,54) = 92.124$, $p < .001$.

The category coordinates showed a semantic priming effect, $\mathrm{F}_1(1,19) = 1754.725$, $p < .001$, $\mathrm{F}_2(1,52) = 104.371$, $p < 0.001$, and an associative priming effect, $\mathrm{F}_1(1,19) = 44.774$, $p < 0.001$, $\mathrm{F}_2(1,52) = 4.647$, $p < .05$. No associative boost appeared in either analysis due to the surprisingly large priming effect for non-associated artifacts.

Semantic priming was present in the functional relations, $\mathrm{F}_1(1,19) = 892.731$, $p < .001$, $\mathrm{F}_2(1,52) = 96.693$, $p < .001$. Associative priming was significant across subjects, $\mathrm{F}_1(1,19) = 47.997$, $p < .001$, and marginally significant in the items analysis, $\mathrm{F}_2(1,52) = 3.403$, $p = .07$, The associative boost was significant for subjects, $\mathrm{F}_1(1,19) = 51.055$, $p < .001$, and approached significance in the items analysis, $\mathrm{F}_2(1,52) = 3.168$, $p = 0.081$.

Separate analyses for each semantic subtype are reported elsewhere [12].

3.8 Discussion

Table 2 shows that low-dimensional simulation gave results very similar to those found in the original experiment, and in its replication in high-dimensions. As is typically the case in dimensionality reduction, related items become more similar. This can be seen in Table 2 where priming effects are much larger, whereas the unrelated baseline is essentially unchanged. The reduction also brings out previously weak trends in the data, e.g. the fact that the non-associated semantic priming effects is much stronger for category coordinates, and that instrument relations do not require association to prime strongly.

4. Conclusion

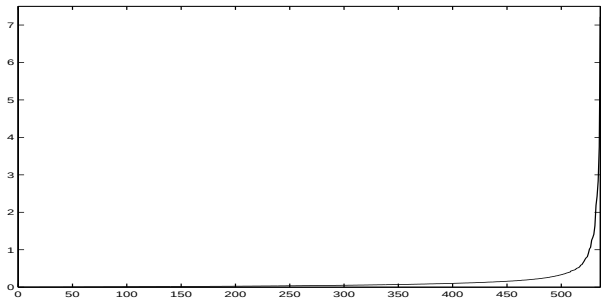

Figure 1: Eigenvalues of the covariance matrix for the high-dimensional semantic space vectors, ordered by size.

Experiment 2 also suggests that the the intrinsic dimensionality of the semantic space data is quite low. Another complementary way to see this is to look at linear measures of the variance structure. Figure 1 shows the eigenvalues of the covariance matrix for the high-dimensional data, sorted by size. It is clear from the figure that the majority of the data variance extends in only a few directions. The first handful of values contain more than 80% of the total variance. Another way to understand this is to consider linear reconstructions of the data: only a handful of real numbers representing data

projections onto the principal eigenvectors would be necessary to reconstruct this data to 80% accuracy.

Looking at orthogonal directions of variance is a useful baseline for understanding the success of the GTM because the intrinsic dimensionality of the data can only be smaller than a linear estimate would suggest. On the other hand the procedure is only approximate since the interpretation of eigenvalue structure in terms of variance component holds only for jointly Normally distributed data. This assumption is unlikely to hold exactly for semantic space vector elements.

In any case it is interesting to compare this to Landauer and Dumais' claim that several hundred dimensions are necessary for semantic space. It is possible that the tasks used in that work require significantly different dimensionality spaces than for even fairly detailed priming studies. Replicating the Landauer and Dumais' tasks in the current framework is current work. But this is clearly not the case for this data. We have also shown elsewhere that many other priming results can be captured in very low-dimensional models [12]. These studies support the claim that the dimensionality of human semantic space may be very low indeed.

Acknowledgements

Thanks to Daniel Dennett for supporting this work, and to two anonymous reviewers for helpful comments.

References

1. Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.
2. Bishop, C. M., Svensén, M., and Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235.
3. Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25):211–257.
4. Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*.
5. Goodhill, G. J. and Willshaw, D. J. (1994). Elastic net model of ocular dominance: Overall stripe pattern and monocular deprivation. *Neural Computation*, 6:615–621.
6. Kaski, S. (1989). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of the International Joint Conference on Neural Networks*, pages 413–418.
7. Kohonen, T. (1995). *Self-organizing maps*. Springer, Berlin.
8. Lafferty, J., Sleator, D., and Temperley, D. (1992). Grammatical trigrams: A probabilistic model of link grammar. Technical report, CMU School of Computer Science.
9. Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, (104):211–240.

10. Lowe, W. (1997a). Meaning and the mental lexicon. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 1092–1097, San Francisco. Morgan Kaufmann.
11. Lowe, W. (1997b). Semantic representation and priming in a self-organizing lexicon. In Bullinaria, J. A., Glasspool, D. W., and Houghton, G., editors, *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 227–239, London. Springer-Verlag.
12. Lowe, W. (2000). *Topographic Maps of Semantic Space*. PhD thesis, Institute for Adaptive and Neural Computation, Division of Informatics, Edinburgh University.
13. Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.
14. Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665. Mahwah, NJ: Lawrence Erlbaum Associates.
15. McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, pages 675–680, New Jersey. Lawrence Erlbaum Associates.
16. Moss, H. E., Ostrin, R. K., Tyler, L. K., and Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (21):863–883.
17. Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, (61):241–254.
18. Scholtes, J. C. (1991). Neural nets and their relevance for information retrieval. Technical Report CL-91-02, University of Amsterdam, Institute for Language, Logic and Information, Department of Computational Linguistics.