In: R.F. French & J.P. Sougne (Eds), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop,* 273-282. London: Springer.

Learning Lexical Properties from Word Usage Patterns: Which Context Words Should be Used?

Joseph P. Levy [1] & John A. Bullinaria [2]

[1] School of Social Sciences, University of Greenwich, UK

[2] Department of Psychology, University of Reading, UK

Abstract

Several recent papers have described how lexical properties of words can be captured by simple measurements of which other words tend to occur close to them. At a practical level, word co-occurrence statistics are used to generate high dimensional vector space representations and appropriate distance metrics are defined on those spaces. The resulting co-occurrence vectors have been used to account for phenomena ranging from semantic priming to vocabulary acquisition. We have developed a simple and highly efficient system for computing useful word co-occurrence statistics, along with a number of criteria for optimizing and validating the resulting representations. Other workers have advocated various methods for reducing the number of dimensions in the co-occurrence vectors. Lund & Burgess [10] have suggested using only the most variant components; Landauer & Dumais [5] stress that to be of explanatory value the dimensionality of the co-occurrence vectors must be reduced to around 300 using singular value decomposition, a procedure related to principal components analysis; and Lowe & McDonald [8] have used a statistical reliability criterion. We have used a simpler framework that orders and truncates the dimensions according to their word frequency. Here we compare how the different methods perform for two evaluation criteria and briefly discuss the consequences of the different methodologies for work within cognitive or neural computation.

1. Introduction

Distributional statistics are measurements of simple patterns within data, usually based on patterns of co-occurrence. These methods have been used frequently in psychological models of language phenomena including phonology, morphology and word meaning [13]. In many of these models, neural network learning algorithms are employed to measure the co-occurrence statistics. In this paper we report on a study in which the statistics are measured directly. However, the results remain relevant to connectionist and other intelligent systems techniques that involve the use of distributional patterns.

Recently, several groups have made claims that aspects of lexical semantics can be captured by looking at patterns of lexical co-occurrence [9, 10, 5, 8]. We have been looking carefully at the techniques used to collect and make use of these statistics and have shown that varying the precise details of the computational procedures employed can make large differences to the way that the measurements

can account for language based data [12, 6].

These techniques have numerous different applications within cognitive science and language technology. Within psychology, the fact that information is there to be extracted counters some "poverty of the stimulus" arguments against the possibility of learning language structure. The representations can also be used in cognitive models and in theories of learning. The techniques have already been successfully applied in language technology to achieve semantic disambiguation and document retrieval [3, 14]. It will be interesting to see whether the interdisciplinary study of exactly how these techniques work will reap benefits for both cognitive science and practical information technologies.

The idea of using distributional statistics to examine aspects of lexical semantics comes from the intuition that some aspects of word meaning may be deduced from the way in which the given word is used in relation to other words. Word usage can be measured by looking at the patterns of co-occurrence between the target word of interest and large numbers of other words in a corpus of real language use. Of course, this technique leaves out several important aspects of word meaning and language use, e.g. reference to objects and events in the world; the influence of general knowledge; speech cues such as stress and prosody. However, recent work has shown that differences in simple patterns of word usage do appear to reflect differences in word meanings.

There have been two broad classes of methods used to measure lexical semantic similarity as distances between vectors of co-occurrence counts (e.g. see [11] p 286): *document space* and *word space*. A notable example of work using document space is due to Landauer & Dumais [5] who used a technique that was developed for a practical use – information retrieval. They call their method "Latent Semantic Analysis" or LSA and stress the importance of dimensionality reduction. Using an encyclopaedia designed for children with 4.6 million words and 30,473 articles (with long articles truncated to the first 2,000 characters) they generated vectors for each word with components corresponding to the number of occurrences of the word in each article. They then transformed their vectors using an entropic measure and extracted the 300 most important dimensions using "singular value decomposition" (SVD), a procedure related to principal component analysis. They showed that the learning rate of their model mirrors the pattern of vocabulary acquisition of children and how a child can induce the rough meaning of a previously unseen word from its present context and a knowledge of past word co-occurrences.

Methods that employ word space (e.g. [14, 4, 10, 6]) rely on a more fine-grained counting scheme whereby each component of the co-occurrence vector for each particular (target) word corresponds to a another particular (context) word. The value of each component is an appropriately transformed count of how often the corresponding context word appears close to (i.e. within a particular sized window around) the target word of interest.

2. Our Vector Generation Approach

We [12, 6, 7] use a similar word space method to that used by Lund & Burgess [10, 9]. The set of word co-occurrence vectors forms a matrix consisting of columns for each target word type and rows which represent the counts of how often each

The lorry *driver* swerved on the road. As well as causing *pollution*, a lorry also has large *wheels*. A lorry requires *diesel* to work. A lorry might carry *sweet* apples and bananas. Bananas are easier to *peel* than apples but apples have nicer *trees*. Bananas are cheaper than apples in a *shop*.

	lorry	apples	bananas
sweet	1	1	2
trees	0	2	2
shop	0	0	1
eat	0	0	0
peel	0	2	2
driver	1	0	0
road	1	0	0
diesel	2	0	0
pollution	1	0	0
wheels	2	0	0

Table 1: Simple example of collecting the word co-occurrence statistics.

context word type occurs within a small window around the target word. Our counts are usually derived from the textual component of the *British National Corpus* (BNC) [1] which consists of 90 million words from a wide variety of sources, though we have also explored using vectors from an 168 million word corpus of USENET newsgroup text. Whatever the source, we create what is effectively a simple probability distribution by normalising each raw count by word frequency and window size to get the conditional probability of each context word type appearing around each target word.

The manner in which we measure the word co-occurrence statistics can be seen more clearly in the simple example shown in Table 1. The block of text is our corpus, the underlined words are our target words, and the *italicised* words are the context words. We count the number of times each context word appears in a window of plus or minus five words around the target words. This gives frequency counts that describe the typical context of each of the target words in terms of the context words they co-occur with. In this way we derive ten dimensional vectors for the three target words from the corpus of 50 words. This example serves to show how the raw statistics are counted before being normalised. The co-occurrence vectors (columns in the table) become probability distributions when the raw counts are divided by the overall frequency of occurrence of the target word with an adjustment made for window size. In practice, of course, we would use a very much larger number of target and context word types and a considerably larger corpus.

If the vector components are ordered according to the total number of occurrences in the whole corpus of the corresponding context words, we then have a straightforward procedure for reducing the dimensionality of the vectors by removing the lowest frequency components.

Various claims have been made in the literature about better ways of restricting or reducing the numbers of dimensions used. Some of these methods are merely practical considerations for computational convenience, and some may reflect the statistical nature of the language data, but others may have important implications for how such statistics are used cognitively or neurally. This paper will concentrate on comparing which context word dimensions really are most useful for this kind of work.

3. Other Previous Work

3.1 Finch & Chater

Finch & Chater [4] explored how co-occurrence vectors might serve as a basis for inducing syntactic categories. They took a 40 million word corpus of USENET newsgroup text and used the 150 most common words in the corpus as context words with a window of two words either side of each target word. They analysed the data from the 1000 most frequent words and found that a simple vector correlation technique revealed a considerable amount of information about syntactic categories. They found that cluster analysis dendrograms could be interpreted as a hierarchy of syntactic categories that is remarkably close to a standard linguistic taxonomy and included structure right up to phrasal categories.

Although this work concentrated on inducing syntactic regularities, they also found that some of their clusters exhibited semantic regularities. The most common 150 words in a corpus of English are likely to be mostly closed class or grammatical function words such as determiners (e.g. "the", "a"), conjunction, prepositions etc. It is not terribly surprising that the syntactic category of a word is found to be related to its pattern of co-occurrence with function words. What is more surprising is that function word co-occurrence gave some flavour of semantic properties. The use of closed class word co-occurrence patterns to induce measures of semantic similarity will be examined further below.

3.2 Lund & Burgess

Lund & Burgess have published several recent papers using co-occurrence statistics within a framework that they call "Hyperspace Approximation to Language" or HAL. In one particular study [10], they used a 160 million word corpus of USENET newsgroup text which they claim to be a source that gives them natural conversational language. They used a weighted window of size 10 to produce their word co-occurrence counts. The Euclidean distances between words in the high dimensional space were then used to predict the degree of priming of one word with the other in a lexical decision task.

They found that their results were unchanged by including more than the first 200 most variant context words. Their use of only the most variant words appears to be made on the grounds of computational convenience rather than a claim for limiting the scope of context words that might be used in the brain.

3.3 Lowe & McDonald

Lowe & McDonald [8] have described the use of co-occurrence vectors to model mediated priming. They collected their word co-occurrence statistics using a window size of 10 and took a positive log-odds ratio as a measurement of lexical association. They chose the context word dimensions to use in their model by selecting only those that were most "reliable". These were chosen conservatively using an ANOVA to judge how consistent the co-occurrence patterns of the context words were across different sub-corpora. Using a rather conservative criterion, this

method yielded 536 context words. Before they measured the reliability they ruled out a "stop-list" of 571 words including closed class words and other mostly very common words which are usually seen as uninformative in the information retrieval literature.

4. Comparing Different Context Word Dimensions

In this paper we shall concentrate on studying the choice of context word sets that form the basis for the acquisition of the co-occurrence vectors. We begin by comparing how the different strategies for choosing the context word dimensions performed on an evaluation measure based on the synonym portion of the "Test of English as a Foreign Language (TOEFL)" task used by Landauer & Dumais [5]. This consisted of a set of 80 test words and the task was to choose the word most closely related in meaning to each test word from a set of four alternatives. Tom Landauer kindly provided us with the materials for this task. Although this test was originally used to demonstrate the utility of their LSA framework, we shall compare the different context word choices within our (word space) framework.

The LSA program scored around 64% by using a strategy of choosing the word with the largest cosine (i.e. smallest angle) between it and the target. This score is comparable to the average score by applicants to U.S. colleges from non-English speaking countries and is apparently high enough to fulfil that part of the admission requirements for many U.S. universities.

Before we can compare the various different context word sets, we need to make various other design choices, such window type and size, distance measure, corpus size, and so on. We have already considered these decisions in some detail elsewhere [12, 6, 7], and from these studies we choose what appears to be the best overall set-up for this task. We consequently used a window size of two words to the left and two to the right to collect co-occurrence statistics from the 90 million words of the written component of the BNC. We compared word vectors p and q using the Hellinger distance measure

$$H(p,q) = \sum_i (\sqrt{p_i} - \sqrt{q_i})^2$$

which is one of a sequence of information theoretic measures (including the Kullback-Leibler divergence) that is appropriate for comparing probability distributions [15].

4.1 Choosing the Context Word Sets

Figure 1 compares the performance on the TOEFL task within the methodological framework outlined above for four different methods of choosing and ordering the context word dimensions:

Frequency – ordering the context words by their frequency of occurrence in the BNC. We stop after 8192 words by which point, for most of the evaluation measures we have used in the past, performance has levelled off.

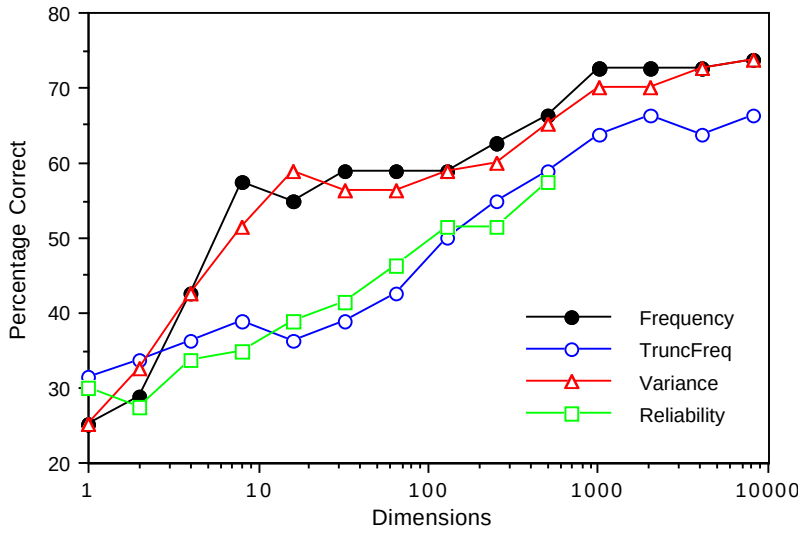

Figure 1: Performance on TOEFL task against number of dimensions.

TruncFreq – again ordering by frequency but with the most frequent 147 words (i.e. words with counts over 50,000) removed. This is to evaluate claims about the utility of the most frequent (usually closed class) context words.

Variance – ordering the context words by the variance of their components across all the target words in the corpus. This is the method used by Lund & Burgess although they used a large (weighted) window and a different corpus.

Reliability – the most reliable 536 words (ordered by frequency) kindly provided by Will Lowe and Scott McDonald from their study [8].

We can see from Figure 1 that there is a clear superiority for the ***Frequency*** and ***Variance*** approaches over the other two, particularly for dimensions less than a few hundred. The ***Reliability*** method does about as well as the ***TruncFreq*** method up to its limit of 536 dimensions where it achieves about 9% less than the simple ***Frequency*** method.

We repeated the above experiment for 19 equal sized non-overlapping sub-corpora of 4.6 million words within the written component of the BNC. These are roughly the same size as the corpus used by Landauer & Dumais [5] and may be closer to the actual number of words actually read by a person during their school education. The general pattern of results was similar but uniformly lower than those for the full 90 million word corpus.

To check the extent to which this pattern of results was specific to the TOEFL task, we repeated the experiment using a second evaluation measure that we have developed and tested in the past, namely a simple semantic categoriser with a large number of candidate categories [6]. For this task, we first chose 10 of the highest frequency members from each of 53 of the Battig & Montague [2] semantic categories. Then for each category we computed the centroid (i.e. the geometric

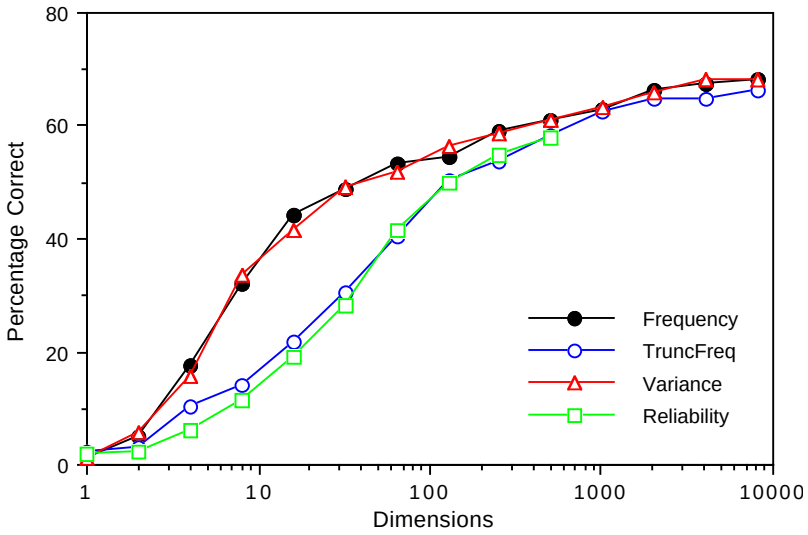

Figure 2: Semantic categorisation scores against number of dimensions.

mean) of the co-occurrence vectors of its 10 members. These are the points in the vector space about which we would expect the category members to cluster. The performance on this task is the percentage of the 530 vectors that are correctly classified in the sense of being closest to the correct one of the 53 centroids. (In practice we need to avoid the bias caused by having each vector used in the definition of its own centroid by excluding it, leaving in each case the 53 centroids defined by the remaining 529 vectors.)

Figure 2 shows the outcome of this second study. We again determined vector closeness with the Hellinger distance measure, but used a larger window size of four words to each side, as we have previously found these to be optimal for this task [6]. We see that the results are broadly in line those for the TOEFL task. ***Frequency*** and ***Variance*** again produced the best scores, but their advantage over ***Reliability*** and ***TruncFreq*** is reduced above around 500 components.

4.2 Frequency versus Variance

We can understand why the ***Frequency*** and ***Variance*** results are so similar from Figure 3. We see that the mean values of each component across the targets words (which are clearly proportional to component word frequencies) are highly correlated with the variances. Truncating the co-occurrence vectors according to one will clearly have a similar effect to truncating according to the other.

4.3 Open and Closed Class Words

In both Figures 1 and 2 it is notable that the results are good but not optimal for the first 100 or 200 dimensions. ***Frequency*** and ***Variance*** do best over this range despite the fact that for both these methods (and almost exclusively for ***Frequency***)

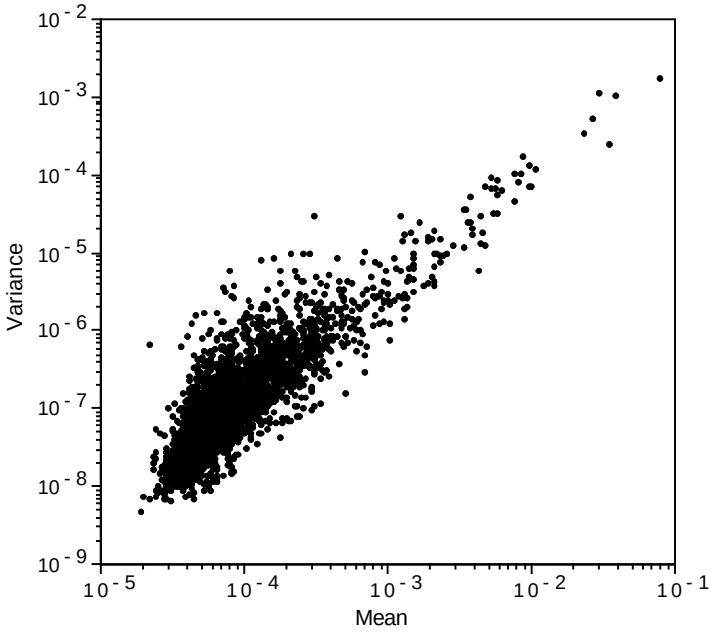

Figure 3: Mean and variance of vector components across target words.

these context word dimensions correspond to closed class or function words. It is sometimes assumed that these words are so frequent and unvarying in their patterns of co-occurrence that they would not give much information about the semantic usage of a word. Lowe & McDonald [8] follow the practice used in the information retrieval literature and exclude a "stop list" of closed class and other presumed unimportant words from consideration as context dimensions. It appears from the results presented here that these words *are* useful for the tasks we have examined and that their method for choosing context word dimensions suffers by excluding them from consideration. In fact, we have found that if the most frequent 147 words from the corpus (which are mostly closed class) are added to the 536 words of the ***Reliable*** set, then results are boosted significantly.

4.4 Dimensionality reduction

Using their document space based approach, Landauer & Dumais [5] found that as long as they chose the first 300 principal components of their dimensionally reduced matrix they achieved good results on the TOEFL test (around 64%). However, the performance on the TOEFL test was poor if too few or too many dimensions were used. They went on to describe a model of vocabulary acquisition in children which accounts for their extraordinarily high rate of word learning. Their paper stresses the importance of using an optimal dimensionality of data derived from the co-occurrence statistics.

We have shown here and elsewhere [6] how, using a larger corpus and our word space based method, we can achieve very good results on the TOEFL test without

any form of dimensionality reduction, apart from ignoring the very low frequency words as a matter of convenience. We have also presented [7] some preliminary results on using singular value decomposition on our co-occurrence vectors. There seemed to be no improvement demonstrated over the original vectors, even when we used our small corpora of 4.6 million words. Clearly, dimensionality reduction is useful in some cases, but only for certain ways of using the co-occurrence statistics. For our word space based method using a small window size and an information theoretic distance measure, dimensionality reduction appears to be unnecessary. Clearly this feature warrants further investigation in the future.

5. Discussion

It is clear that all the methods we have discussed and compared in this paper extract useful and interesting statistical regularities from corpus-based co-occurrence counts. We believe that the methodological claims that are made in this field should always be tested using different data and different parameters to determine how general they really are. Our approach has been to start with as simple a procedure as possible and to tune the various parameters (such as window type and size) by observing empirically how well the different parameter combinations perform under various different evaluation tasks (such as the TOEFL test and our semantic categorisation task). We have demonstrated here that simply ordering the context word dimensions in order of their frequency and using as many of them as possible or practical produces good results. Closed class or function words appear to provide useful contributions to the co-occurrence statistics. Ordering the context words by variance or reliability appears to confer no advantages within our framework and the computationally expensive procedure of singular value decompositions seems not to increase performance on our evaluation tasks.

Choosing between methodological variations is often a matter of computational convenience, or corresponds to a claim or assumption about the statistical language data, rather than a statement about how cognitive and neural systems might realise and utilise such data. However, the use of closed class words probably does have implications for cognitive models because there are claims in the psycholinguistic literature that open and closed class words are processed differently in the brain. We have found no evidence that this has to be true because including closed class words does seem to improve performance in the corpus based studies discussed in this paper. Moreover, there are ways in which connectionist and neural systems might instantiate dimensionality reduction, but again we have found no evidence that this has to take place to make good use of word co-occurrence statistics.

References

1. Aston, G. & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.
2. Battig, W.F. & Montague, W.E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, *80*, 1-45.

3. Dagan, I., Marcus, S. & Markovitch, S. (1993). Contextual word similarity and estimation from sparse data. in *Proceedings of the 31st Annual Meeting of the ACL*, 164-171.
4. Finch, S. & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, 820-825.
5. Landauer, T. & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104(2),* 211-240.
6. Levy, J.P., Bullinaria, J.A. & Patel, M. (1998). Explorations in the derivation of word co-occurrence statistics. *South Pacific Journal of Psychology*, *10(1),* 99-111.
7. Levy, J.P. & Bullinaria, J.A. (1999). The emergence of semantic representations from language usage**.** Paper given at the *EPSRC Workshop on Self-Organising Systems - Future Prospects for Computing*, UMIST, October 1999.
8. Lowe, W. & McDonald, S. (2000). The direct route: Mediated priming in semantic space. *In Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.*
9. Lund, K., Burgess, C. & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedin*gs of the 17th Annual Meeting *of the Cognitive Science Society*, 660-665.
10. Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence, *Behavior Research Methods, Instruments, & Computers*, *28(2),* 203-208.
11. Manning, C.D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
12. Patel, M., Bullinaria, J.A. & Levy, J.P. (1998). Extracting Semantic Representations from Large Text Corpora. In Bullinaria, J.A., Glasspool, D.W. & Houghton, G. (eds), *4th Neural Computation and Psychology Workshop, London, 9-11 April 1997: Connectionist Representations*, 199-212. London: Springer-Verlag.
13. Redington, M. & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences*, *1 (7),* 273-281.
14. Schütze, H. (1993). Word Space. In S.J. Hanson, J.D. Cowan & C.L. Giles (Eds.) *Advances in Neural Information Processing Systems 5,* 895-902. San Mateo, CA: Morgan Kauffmann.
15. Zhu, H. (1997). Bayesian Geometric Theory of Learning Algorithms. In: *Proceedings of the International Conference on Neural Networks (ICNN'97),* Vol. 2, 1041-1044.