In E.J. Davelaar (Ed.), Connectionist Models of Neurocognition and Emergent Behavior: From Theory to Applications, 292-308. Singapore: World Scientific, 2012.

USING ENRICHED SEMANTIC REPRESENTATIONS IN PREDICTIONS OF HUMAN BRAIN ACTIVITY

JOSEPH P. LEVY

Department of Psychology, Roehampton University, Holybourne Avenue,
London, SW15 4JD, UK

JOHN A. BULLINARIA

School of Computer Science, University of Birmingham, Edgbaston,
Birmingham, B15 2TT, UK

There have been many different theoretical proposals for ways of representing word meaning in a distributed fashion. We ourselves have put forward a framework for expressing aspects of lexical semantics in terms of patterns of word co-occurrences measured in large linguistic corpora. Recent advances in the modelling of fMRI measures of brain activity have started to examine patterns of activation across the cortex rather than averaging activity across a sub-volume. Mitchell et al. [11] have shown that simple linear models can successfully predict fMRI data from patterns of word co-occurrence for a task where participants mentally generate properties for presented word-picture pairs. Using their MRI data, we replicate their models and extend them to use our independently optimised co-occurrence patterns to demonstrate that enriched representations of word/concept meaning produce significantly better predictions of brain activity. We also explore several aspects of the parameter space underlying the supervised learning techniques used in these models.

1. Introduction

There have been many suggestions for methods of representing word or concept meaning in terms of a distributed pattern of feature values [e.g., 9,14]. Some of these reflect linguistic intuitions of participants (or experimenters/modellers) and others measure the distributions of words in language corpora. One of the challenges in the field is to test whether a particular scheme for semantic representation can explain or predict human behaviour better than another scheme. Recent work by Mitchell, Shinkareva, Carlson, Malave, Mason & Just [11] has shown that the co-occurrence statistics for a small hand-picked set of verbs can be used to predict functional brain imaging data at a level well above chance. In this paper, using the brain imaging data they have generously made public, we compare their results with the performance of their method using our own co-occurrence statistics that have been independently optimised.

Magnetic resonance imaging (MRI) is a form of spectroscopy that can produce high-resolution 3-dimensional images of materials including the anatomy of the brain. This technique can be extended to produce images that reflect blood deoxygenation in the brain where the strength of the "blood oxygen level dependant" or BOLD signal is measured for subvolumes or "voxels" of brain tissue of around 3-5mm-cubed and assumed to correlate with neural activation. The functional (fMRI) signal can usually only be measured relative to a resting or contrasting state and is small and noisy. The usual and highly successful method of analysing fMRI data is the so-called mass-univariate approach where a map is produced of the inferential statistic for each voxel produced by a linear statistical model of the predictors in an experiment. With appropriate adjustments for multiple comparisons, this approach allows a map to be produced showing areas of the brain sensitive to the experimental prediction or for a prediction that a particular area is sensitive to the experimental contrast to be statistically tested. However, the method ignores the possibility that what may be interesting in the data is the *pattern* of activation across individual voxels rather than the level of activation of a voxel or its mean across voxels. Mitchell et al's work is an example of this increasingly popular use of BOLD signal voxel patterns.

As reviewed by Naselaris et al. [12], these pattern analysis techniques have the potential for wide application in cognitive neuroscience. Here, we explore the possible ways of analysing the encoding of lexical or conceptual meaning in the brain. Specifically, we test the efficiency of different linear mappings from putative semantic representations to patterns of brain activation. In general, this allows us to identify distinct brain areas whose activity can be predicted by a particular representational scheme, but here we concentrate on comparing the encoding accuracy of linear computational models built using two different kinds of lexical co-occurrence vectors. The eventual aim is to test the validity of different ways in which neural computational models represent word or conceptual meaning by using measures derived from brain physiology.

In the following section we describe the Mitchell et al. model in more detail, and discuss the key performance measures we employ. Then we outline our own approach to corpus derived semantic representations, and present the parameter optimisation process involved in using them effectively. The comparative results for the key tasks in the Mitchell et al. study are then presented for both regularised and unregularised models, and the effect of varying the training data set size is explored. The paper ends with our conclusions and some discussion.

2. Description of the Computational/Statistical Model

Mitchell et al. asked participants to mentally generate properties for 60 previously studied simultaneously presented word-picture pairs of concrete entities, such as vehicles or animals (e.g., *cat*, *cow*, *train*, *airplane*), whilst lying in an fMRI scanner. Each word-picture pair was presented on 6 different occasions. The scanner data yielded vectors of BOLD activation (relative to a baseline) across the voxels for all the grey matter in the brain being scanned. Feature selection was achieved by choosing the 500 voxels whose values were most *stable* across the 6 repeats for each word (measured by averaging the pairwise correlations between the 60-item vectors produced for each of the presentations of the 60 words for each voxel). The BOLD signal values were then averaged across the repeats for each word and normalised to produce vectors of 500 continuous values for each of the 60 words.

Mitchell et al. hand-picked 25 verbs whose co-occurrences with the nouns in question could be expected to distinguish the patterns of usage and hence meaning of the nouns. The co-occurrences were measured in the 1 trillion word Google corpus to yield an 25-dimensional vector representing word/concept meaning.

To train a model to predict brain activation from word meaning, 58 of the 60 words were used to fit a linear regression model predicting each of the 500 most stable voxels for those 58 words from the 25 feature co-occurrence vectors. The models were then tested on their ability to predict the activation of the two held-out words. The model was deemed to have a correct prediction if the sum of the distances between measured and predicted brain activation was smaller for the correct mapping of input vector to output vector compared to the incorrect mapping. The process was then repeated until all the 1770 combinations of training set and test set had been trained and tested and the success of the model expressed as the mean binary scores.

The exercise produced highly statistically significant results as measured against a distribution of randomly permuted models. As such, it is an important demonstration that a distributed pattern known to reflect some of the properties of a stimulus (the 25 co-occurrence features) can be used to make statistical predictions of putative measurements of brain activity.

Mitchell et al. speculate that a richer representation of word meaning might yield better results in their models. They have generously made their data available [http://www.cs.cmu.edu/~tom/science2008/index.html] and we describe here how we have replicated their work and extended it to use our semantic representations. This advances their work by using vectors that can be shown to have better semantic distinctiveness and are general to all words of a reasonable frequency in the corpus we used to generate the vectors. However, these advantages for the input features of the model are offset by the practical considerations of increasing the dimensionality of the input from 25 to tens of thousands, which results in much slower computation of the linear models and the need for regularisation to prevent overfitting.

Following Mitchell et al., we used MATLAB to compute multiple linear regression models to predict each output voxel value from the values of all input features. This is equivalent to computing a multiple regression for each output voxel consisting of a linear combination of predictors from the input feature values. These computations are conveniently expressed as the minimisation of the sum-squared output error E of the model over a set of training items i:

$$E = \sum_i |m_i - v_i|^2 \qquad , \qquad m_i = Wf_i$$

where f_i is the vector of features, v_i is the vector of voxels, and m_i is the vector of model outputs, for word i. The matrix W of model weights/coefficients can be computed easily using standard matrix pseudoinversion techniques. Mitchell et al. report results from models that used their 25 input features and the 500 most stable voxels across the training set. We attempted to exactly replicate their model and also investigated the effects of varying the number of voxels used and the dimensionality of our own input vectors when they were used.

Regularisation techniques help avoid the overfitting of a model when the ratio of predictors to data points is high. The standard approach is to add a term proportional to the sum of the squares of the model coefficients W to the sum-squared error term E that is being minimised by least-squares learning. This penalises model complexity and helps avoid the fitting of the model to noise rather than true data, and is equivalent to ridge regression. We report later on the optimisation of the parameter that multiplies the regularisation term.

The aim is to test generalisation, i.e. how well the model outputs m_i match the actual voxel patterns v_i for unseen input words i. Following Mitchell et al., we measure this similarity using the cosine $\cos(m_i, v_i)$ between the relevant vectors. For small data sets (such as the 60 words here) a cross-validation approach is appropriate: withhold each possible pair of words (i, j) from training and for each withheld word i determine whether the model output is more similar to the corresponding voxel pattern or the voxel pattern of the other withheld word, i.e. whether $\cos(m_i, v_i) > \cos(m_i, v_j)$. That gives 59 tests for each word, and the model performance is the average number of correct matches over the 3540 tests (which we shall call *Perf*). This *Perf* is the cross-validated estimate of the average probability p that the model output for a given word is closer to the correct word target output than that of another word. It can therefore be used to estimate the expected average performance on harder tasks, such as leaving N words out of training and seeing how often the correct word is closest (i.e. probability p^{N-1}) or where the correct word ranks in closeness among the other N-1 words (i.e. position 1+(1-p)(N-1)). Mitchell et al. actually combined the cosine similarities for each word pair, i.e. tested whether the pair total $\cos(m_i, v_i) + \cos(m_j, v_j) > \cos(m_i, v_j) + \cos(m_j, v_i)$, and took their performance to be the average number of those correct matches over the 1770 word pairs (which we shall call *PairPerf*). If the semantic features were random, both performance measures would be 0.5, and for totally successful models both would be 1.0. For intermediate levels, however, *PairPerf* will generally be higher than *Perf*. Permutation tests show empirically that the 0.05 significance level falls at 0.58 for *Perf* and 0.62 for *PairPerf*.

3. Features used in the modelling

We aim to compare the hand-picked 25 feature set used by Mitchell et al. with our own much larger feature vectors. One way to judge whether it is likely that our feature set will perform better in predicting fMRI data from a task that taps lexical/semantic judgments is to perform an unsupervised cluster analysis on the two sets of vectors for each of the 60 words used by Mitchell et al. The word-picture pairs describe 5 instances each of 12 conceptual categories (e.g., animals, plants, tools, buildings) and so it should be possible to see some of this category structure in the cluster analysis.

The purity P_r of a cluster r is simply the fraction of its members that belong to the most represented class. Then the overall purity P of clustering is the weighted average of the individual cluster purities P_r. Formally

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P_r \qquad , \qquad P_r = \frac{1}{n_r} \max_c \left(n_r^c \right)$$

where n_r and n_r^c are the numbers of words in the relevant clusters and classes, with r labelling the k clusters, and c labelling the classes [15]. Applying the CLUTO Clustering Toolkit [7] with cosine distance and default parameters shows the 25 dimensional features they use to have a purity of 0.47, and the 500 most stable voxels give purities 0.53, 0.42, 0.45, 0.47, 0.45, 0.43, 0.37, 0.33, 0.38 (mean 0.43) for the nine participants.

It is known that optimised corpus-derived semantic representations can achieve perfect purity (1.00) for un-ambiguous concrete nouns [2], so it makes good sense to explore whether using such representations can improve the results in the Mitchell et al. study. We have previously carried out a systematic exploration of how to generate the best semantic representations and found that simply computing vectors of probabilities that words occur next to each other in large corpora, and calculating point-wise mutual information (PMI) [10] leads to vectors that perform well across a range of tasks [1]. It is that type of semantic representation, derived from the two billion word ukWaC corpus [4], that was used to explore clustering ability and achieved perfect purity in some cases [2]. Using such vectors with 10,000 components achieves cluster purity of 0.83 on the 60 Mitchell et al. words. This is still not perfect, and it is clear that such corpus derived vectors never can work well for all words [1, 2, 5, 8], but the improved clustering over the Mitchell et al. features suggests that they are worth testing in the Mitchell et al. model. It might then be expected that, although more computationally expensive, our vectors may be able to predict fMRI data more successfully.

4. Parameter Optimisation

We began our experiments with an exact replication of the Mitchell et al. study, which involved computing the voxel stability values using only the 58 training set items so there could be no issue of non-independence in using the test items during training. However, simply computing the most stable voxels across the entire 60 member training + testing set once for all models, instead of once for each of the 1770 training sets, actually proved to have very little effect on the resulting performance, and allowed massive improvements in computing time, so that approach was followed for the remainder of this study.

We surveyed the different mean performance results for both the *PairPerf* and *Perf* criteria for all 1770 combinations of 58 item training sets and 2 item test sets for a large range of values for the regularisation parameter, the number of frequency-sorted corpus components (for our input features) and the number of stability-sorted voxels. This allowed us to identify near optimal parameter values for both types of input vectors. Each point in the following graphs shows the mean performance over the 9 participants and thus summarises 9 x 1770 = 15,930 trained models.

It is interesting to see how performance changes as the size of the input feature set increases and this is shown in Figure 1 where performance is plotted against number of input components (ordered by associated word frequency in the ukWaC corpus) for different numbers of stability-sorted voxels. Performance peaks at around 10,000 components for both the *PairPerf* and *Perf* criteria. This

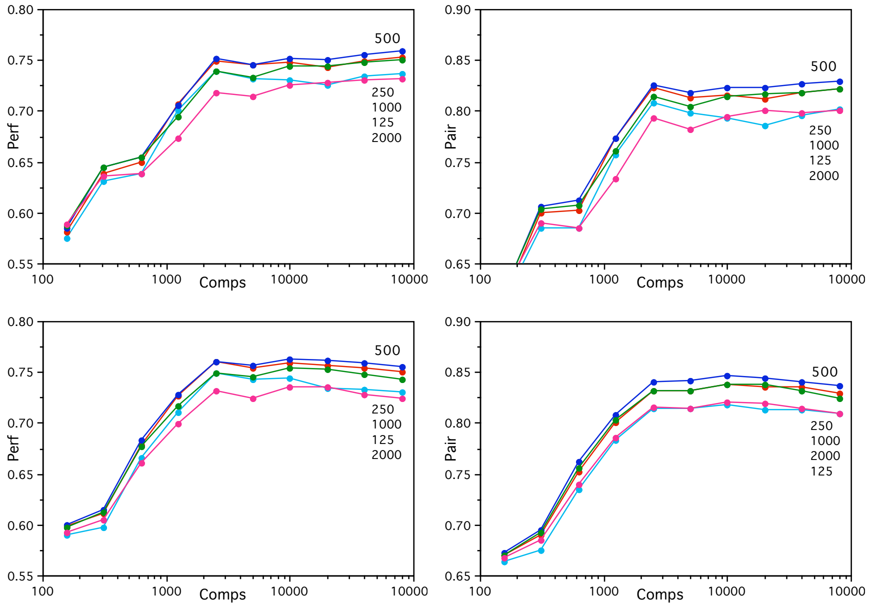

Figure 1. Performance using our vectors as a function of number of corpus components: *Perf* (left) and *PairPerf* (right), regularization parameter 0 (upper) and 10,000 (lower). Individual lines correspond to different numbers of voxels as indicated by the line labels.

pattern is familiar for this kind of corpus-derived representation – the increased information from more components is eventually out-weighed by the noise that comes from poorer estimates of co-occurrence probabilities for lower frequency words [1,2]. These graphs also show that for these data, averaged over the 9 participants, 500 voxels is consistently the best performing number of voxels.

Figure 2 shows that for our 10,000 component input vectors, performance peaks at a regularisation parameter of around 100 for the *Perf* criterion and 300 for the *PairPerf* criterion, and remains fairly level for higher values. It also shows more clearly how the performance falls off for more or fewer than 500 voxels. However, we have observed that for individual participants with the most stable voxels (perhaps simply due to not moving whilst in the scanner as noted by Mitchell et al.), larger numbers of voxels are advantageous and that leads to different optimal regularisation parameters.

For fair comparison, we also checked the optimisation of parameters for the Mitchell et al. study. Figure 3 shows that for the 25 input features used by Mitchell et al., optimal performance was obtained for a regularisation parameter of 1 and 500 voxels. Choosing optimal parameters for such models should

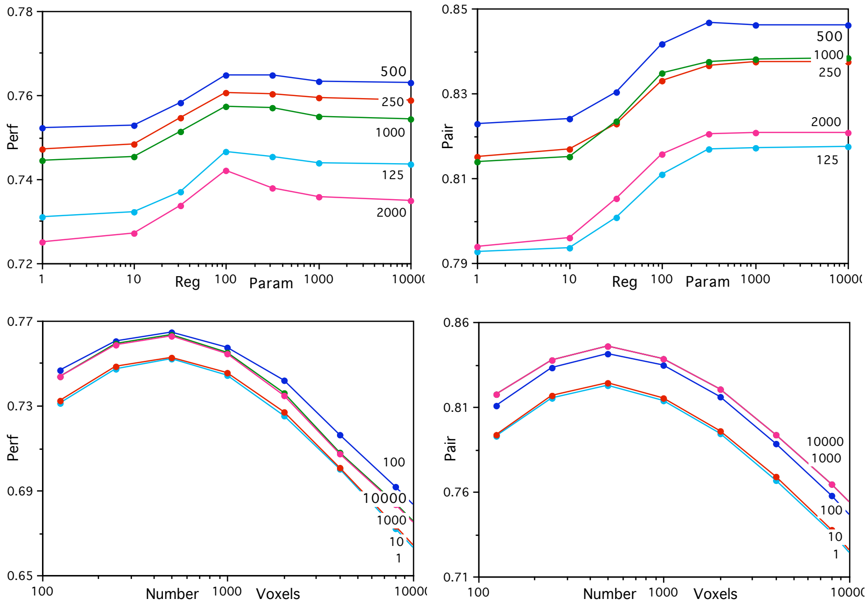

Figure 2. Optimization of performance using our 10000 components corpus vectors: *Perf* (left) and *PairPerf* (right), regularization parameter (upper) and number of voxels (lower). Individual lines correspond to different numbers of voxels (upper) or regularization parameter (lower) as indicated by the line labels. [Note the differing vertical-axis scales.]

really be done using an independent validation set, but with only nine participants of rather variable performance, that was not feasible. It is clear that for these experiments, regularisation makes little difference to the success of the trained models for this task. This might be expected for the low dimensional Mitchell et al. features, and Mitchell et al. did not use regularisation, but we were surprised that our much larger vectors didn't benefit more from its use. For all the comparative studies we used 500 voxels which appears close to optimal on average for both types of input features, and was used in the original Mitchell et al. study. Performance was tested without regularisation, and with regularisation parameter of 1 for the Mitchell et al. features (where there appears to be a consistent peak) and of 10,000 for our 10,000 dimensional corpus features (which is perhaps not optimal, but safely in the flat region of the performance graphs).

Figures 1, 2 and 3 also show how the Mitchell et al. *PairPerf* success measure produces slightly better results than the *Perf* measure across all model parameters. Such a difference is to be expected from a simple consideration of

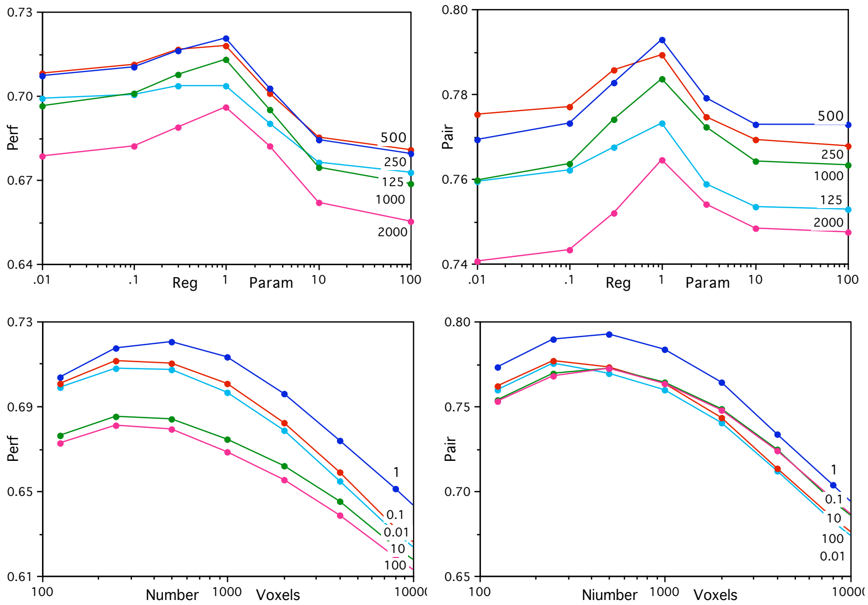

Figure 3. Optimization of performance using the Mitchell et al. features: *Perf* (left) and *PairPerf* (right), regularization parameter (upper) and number of voxels (lower). Individual lines correspond to different numbers of voxels (upper) or regularization parameter (lower).

combining pairs of values from a roughly Gaussian distribution of cosine differences cos(m_i, v_i) – cos(m_i, v_j). However, the nature of the model outputs can lead to the individual cosine differences in the pairs of combined values being anti-correlated, and that can lead to surprising advantages to the *PairPerf* measure. We therefore present both the *Perf* results (that are sure to be free from such artifacts) and *PairPerf* (for comparison purposes).

5. Comparative Results

Now that we have some confidence that we have near optimal values for the main parameters for each of the two different kinds of input features, we can compare performance across the 9 individual participants with and without regularisation. Table 1 shows the results for the individual participants for the *PairPerf* criterion as used by Mitchell et al.: for the 25 dimensional Mitchell et al. feature set, with no regularisation and with regularisation parameter 1, and for our 10,000 component corpus vectors, with no regularisation and with regularisation parameter 10,000. For all cases, 500 voxels were used. Column 2

Table 1. Results for each participant from the original paper by Mitchell et al. [11] (Science), our replication, and for the regularised (with reg) and non-regularised (no reg) models using Mitchell et al. (M. et al.) or Bullinaria & Levy (B & L) features for the *PairPerf* success criterion.

P	Science	Replication	M. et al. no reg	M. et al. with reg	B & L no reg	B & Lwith reg
1	0.83	0.83	0.83	0.84	0.91	0.92
2	0.85	0.85	0.84	0.80	0.78	0.80
3	0.76	0.77	0.78	0.78	0.82	0.85
4	0.78	0.79	0.79	0.82	0.89	0.91
5	0.82	0.82	0.81	0.84	0.78	0.83
6	0.73	0.73	0.71	0.76	0.79	0.83
7	0.78	0.78	0.78	0.78	0.85	0.87
8	0.72	0.72	0.71	0.76	0.70	0.73
9	0.68	0.68	0.69	0.76	0.87	0.88
mean	**0.77**	**0.77**	**0.77**	**0.79**	**0.82**	**0.85**

shows the original results published by Mitchell et al., and column 3 is our replication of what they did. There are small differences that are presumably due to variations in the rounding errors coming from different implementations. Column 4 is the same but with the voxel stability computed just once for the full set of 60 words, rather than for each of the 1770 training sets of 58 words. The differences are slightly larger here, but the mean is the same, which is our justification for using this computationally more efficient approach for the remainder of this study. Column 6 shows the equivalent for our input features. Finally, columns 5 and 7 are the regularised versions corresponding to columns 4 and 6. We tested statistical significance throughout using paired t-tests with one-tailed p-values. In all cases, the statistically significant results are also significant using Wilcoxon non-parametric tests. Our vectors show a modest improvement over those of Mitchell et al. for most participants for both non-regularised ($t(8) = 2.123$, $p < 0.05$) and regularised ($t(8) = 2.877$ $p < 0.05$) conditions. Regularisation appears to make a slight improvement for both feature sets.

Table 2 show the comparison in performance between the input features sets using the *Perf* success criterion. There is again an apparent small advantage for regularisation and statistically significant improvements for our features set over the Mitchell et al. set ($t(8) = 2.111$, $p < 0.05$ for the non-regularised cases and $t(8) = 2.564$, $p < 0.05$ for the regularised cases).

The advantages of our feature vectors are that they did not need to be generated specifically for this task and they are general to a very wide range of

Table 2: Results for each participant for non-regularised and regularised versions of each type of input feature for the *Perf* success criterion.

P	M et al. no reg	M et al. with reg	B & L no reg	B & L with reg
1	0.77	0.76	0.84	0.84
2	0.76	0.73	0.72	0.72
3	0.71	0.71	0.74	0.76
4	0.73	0.76	0.81	0.83
5	0.73	0.75	0.71	0.74
6	0.66	0.70	0.73	0.74
7	0.72	0.70	0.79	0.79
8	0.67	0.69	0.65	0.65
9	0.63	0.69	0.79	0.79
mean	**0.71**	**0.72**	**0.75**	**0.76**

possible words. However, the difference in performance over the Mitchell et al. vectors is rather modest, especially given the computational expense of using 10,000 features rather than 25. To test whether this might be due to a ceiling effect of the fMRI data for this particular task, we progressively increased the difficulty of the task. Figure 4 compares results when the size of the training + test set was successively reduced from 60 words to randomly selected subsets of the original word set of 50, 40, 30, 20 and 10 words (with each case averaged over 20 different subsets). Each data point shows the average and standard error over the nine participants.

The consistency or reliability of the advantage of our input vectors over the Mitchell et al. input vectors appears to increase slightly as the training sets get smaller until a floor effect begins as the set size reaches 10 words. Using paired t-tests to gauge statistical significance of this advantage, it was significant for all training + test set sizes (apart from 10) and the level of significance increased as the training + set size decreased (apart from 10), and hence the difficulty of learning increased giving support to the suggestion that the advantage of our vectors increases as the task increases in difficulty until performance suffers for the regularised models when the set size reaches 10.

6. Further Comparisons

In addition to the main test of their model, described above, Mitchell et al. presented three further tests of its performance that we now apply for our richer representation of semantics.

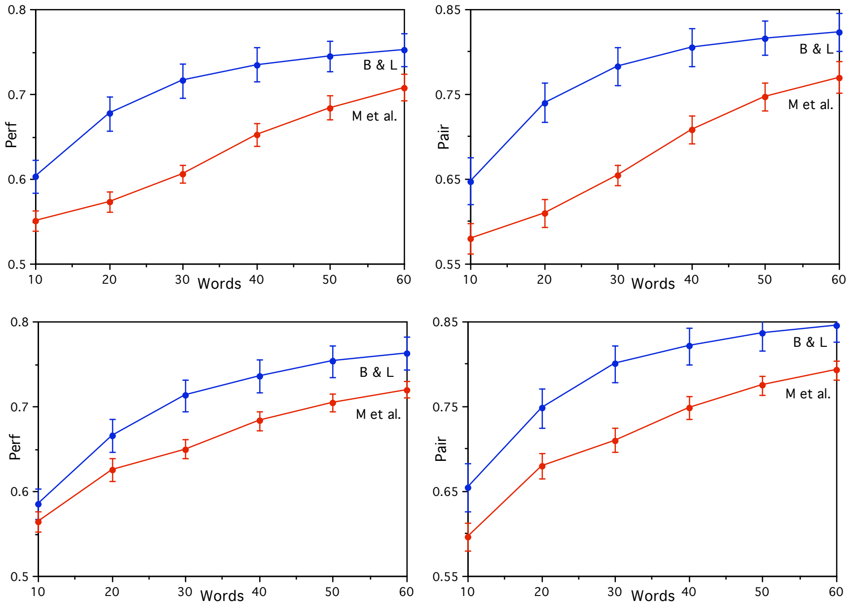

Figure 4. Effect on performance of the number of words in the training data set: *Perf* (left) and *PairPerf* (right), non-regularised (upper) and regularised (lower). Individual lines correspond to using our corpus vectors (B & L) and the Mitchell et al. features (M et al.).

First, since the 60 test words each fall into one of 12 semantic categories, it is instructive to compare the performance for the 120 word pairs that fall within a category, to the full set of word pairs. The within-category task is obviously harder than the full task, and Mitchell et al. did find a performance drop to 0.62. Our replication achieved 0.61 for their features without regularisation and 0.60 with regularisation. For our features, we achieved 0.58 without regularisation and 0.62 with regularisation. The differences between their results and ours was not significant ($t(8)$ = 0.89, p = 0.20 and $t(8)$ = 0.69, p = 0.26). This indicates that the improvements arising from our vectors come primarily from the remaining 1650 word pairs that fall across categories.

Next, we explored how much the presence of semantically related words (i.e. those in the same category) helped the model perform well. The models were retrained for each word pair with all the words from their categories excluded, and the performance recomputed. Mitchell et al. found the mean performance dropped to 0.70. Our replication also achieved 0.70 for their features without regularisation, and 0.74 with regularisation. For our features, we achieved 0.78 without regularisation and 0.79 with regularisation. t-tests showed that our advantages were statistically significant: $t(8) = 2.48$, $p < 0.05$; $t(8) = 2.23$, $p < 0.05$ (both one-tailed).

Finally, we investigated how well the model copes with inputs not from the 60 word test set. For each of 1000 control words (ranked 301 to 1300 in frequency in the corpus) semantic vectors were created and passed through the model trained on 59 of the 60 test words, and the similarity of the withheld word voxel pattern with each of the 1000 control word outputs and 1 withheld word output were ranked. The higher the withheld word ranks (measured as a fraction of the other 1000 words falling below it), the better the models performance. Mitchell et al. achieved a mean performance of 0.72. We did not have the data to attempt replication. For our features, we achieved an improved value of 0.77 without regularisation ($t(8) = 2.1$, $p < 0.05$, one-tailed).

The relatively marginal statistical significance for these last two tasks, despite approximately 5% increases in mean performance, is a reflection of the small number of subjects and the large variations between them. Some subjects consistently perform better with our corpus vectors than with the Mitchell et al. features, and others show the opposite pattern. This could be due to individual differences in the strategy chosen by the participants to perform Mitchell et al.'s relatively unconstrained task being captured better by one set of vectors than another. This is something that is worthy of future investigation, and may well hold the key to refining the whole approach.

7. Conclusions and Discussion

We have been able to closely replicate Mitchell et al.'s results and have shown that regularisation improves them slightly for the original task and more so when the task increases in difficulty. We have also provided results for the single word measure (*Perf*) that we believe is a more generally reliable indicator of performance than Mitchell et al.'s word pair measure (*PairPerf*).

Our richer semantic representations perform better across most of the tasks we have tried. Although the improvements are modest, they are consistent and statistically significant. It appears that the advantage of the much larger and semantically richer feature vectors increases as training set sizes decrease, which may be because the much larger vectors contain enough distributional information to make up for the smaller number of items. It was unsurprising that regularisation was required for our much larger input vectors, but perhaps unexpected that it made so little difference even after the regularisation parameter was optimised. Our average performance peaked when we used 10,000 frequency-ordered components demonstrating that even with a training set of 60 items, the computation of co-occurrence with large numbers of other words was advantageous.

Like Mitchell et al., we found that for this training set, optimal performance overall was achieved by using the 500 most stable voxels. For those individual participants with relatively high average stability values, however, a larger number of voxels was optimal. As with optimising our corpus vectors, there appears to be a trade-off between increased information and increased noise as more components are added.

While we were writing this paper, we discovered other recent work that had also used the Mitchell et al. data set and tested performance for a variety of alternative sets of input features [3, 6, 13]. It is interesting to note that all of these papers report useful results using different input features, but none of them convincingly exceed the performance levels achieved using the Mitchell et al. or our input vectors.

Perhaps the most puzzling aspect of this research is the apparently rather poor results achieved by all of the methods we have described. It is maybe worth speculating on what might be causing such a ceiling effect. It is certainly true that the BOLD signal is always small and subject to large amounts of noise, and that any head movement (which should be reflected in the "stability" measure) will cause problems for a voxel-pattern based approach. However, to train on 58 items and test on the remaining 2 is not an intrinsically difficult task. Mitchell et al.'s input feature set of 25 items performs well above chance and one might expect that our 10,000 element vectors would boost performance more than they did.

It would be instructive to further analyse the degree to which noise in the measurement of voxel BOLD response may be leading to an upper bound in the performance of the learning models reported here. The voxel patterns themselves do not cluster well using the CLUTO algorithm and this may indicate measurement noise. We are currently exploring the degree to which this kind of noise affects the performance of the models by constructing idealised voxel patterns that cluster perfectly and testing how learning performance drops off as noise is added.

It may also be instructive to repeat these modelling methods on a different stimulus set. The particular words chosen by Mitchell et al. are sometimes ambiguous and are always paired with line drawings. It is likely that lexical semantic processes are being confounded with visual ones in ways that differ between the different words. We hope to further investigate these issues using data derived from stimuli that are purely word-based.

Acknowledgments

We are very grateful to Tom Mitchell and Kai-min Kevin Chang for making their data available and for help in replicating their original results. We also thank Eddy Davelaar and an anonymous reviewer for helpful comments on the draft of this paper.

References

1. Bullinaria, J. A. & Levy, J. P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, **39**, 510–26.
2. Bullinaria, J. A. 2008. Semantic categorization using simple word co-occurrence statistics. In: M. Baroni, S. Evert & A. Lenci (Eds), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, 1–8. Hamburg, Germany: ESSLLI.
3. Devereux, B., Kelly, C. & Korhonen, A. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In Proceedings of the First Workshop on Computational Neurolinguistics, Los Angeles, California, USA. Association for Computational Linguistics.
4. Ferraresi, A. 2007. Building a very large corpus of English obtained by web crawling: ukWaC. Masters Thesis, University of Bologna, Italy. Corpus web-site: http://wacky.sslmit.unibo.it/
5. French, R. M. & Labiouse, C. 2002. Four problems with extracting human semantics from large text corpora. *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*, 316–322. Mahwah, NJ: Lawrence Erlbaum Associates.
6. Jelodar, A. B., Alizaseh, M. & Khadevi, S. 2010. WordNet based features for predicting brain activity associated with meanings of nouns. In Proceedings of the First Workshop on Computational Neurolinguistics, Los Angeles, California, USA. Association for Computational Linguistics.
7. Karypis, G. 2003. CLUTO: A Clustering Toolkit (Release 2.1.1). Technical Report: #02-017, Department of Computer Science, University of Minnesota. Web-site: http://glaros.dtc.umn.edu/gkhome/views/cluto.
8. Landauer, T. K. & Dumais, S. T. 1997. A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction and representation of knowledge. *Psychological Review*, **104**, 211–240.
9. Lund, K. & Burgess, C. 1999. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, **28**, 203–208.
10. Manning, C. D. & Schütze, H. 1996. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

11. Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K-M., Malave, V.L., Mason, R.A. & Just, M.A. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, **320**, 1191–1195.
12. Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. in press. Encoding and decoding in fMRI, *Neuroimage*.
13. Pereira, F., Botvinick, M., & Detre, G. 2010. Learning semantic features for fMRI data from definitional text. In Proceedings of the First Workshop on Computational Neurolinguistics, Los Angeles, California, USA. Association for Computational Linguistics.
14. Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R. & Levy, J. P. 2000. Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75(2), 195–231.
15. Zhao, Y. & Karypis, G. 2001. Criterion functions for document clustering: Experiments and Analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota. Available from: http://cs.umn.edu/karypis/publications.