Selecting Corpus-Semantic Models for Neurolinguistic Decoding

Brian Murphy
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
brianmurphy@cmu.edu

Partha Talukdar
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
ppt@cs.cmu.edu

Tom Mitchell
Machine Learning Dept.
Carnegie Mellon University
Pittsburgh, USA
tom.mitchell@cs.cmu.edu

Abstract

Neurosemantics aims to learn the mapping between concepts and the neural activity which they elicit during neuroimaging experiments. Different approaches have been used to represent individual concepts, but current state-of-the-art techniques require extensive manual intervention to scale to arbitrary words and domains. To overcome this challenge, we initiate a systematic comparison of automatically-derived corpus representations, based on various types of textual co-occurrence. We find that dependency parse-based features are the most effective, achieving accuracies similar to the leading semi-manual approaches and higher than any published for a corpus-based model. We also find that simple word features enriched with directional information provide a close-to-optimal solution at much lower computational cost.

1 Introduction

The cognitive plausibility of computational models of word meaning has typically been tested using behavioural benchmarks, such as identification of synonyms among close associates (the TOEFL task for language learners, see e.g. Landauer and Dumais, 1997); emulating elicited judgments of pairwise similarity (such as Rubenstein and Goodenough, 1965); judgments of category membership (e.g. Battig and Montague, 1969); and word priming effects (Lund and Burgess, 1996). Mitchell et al. (2008) introduced a new task in *neurosemantic decoding* – using models of semantics to learn the mapping between concepts and the neural activity which they elicit during neuroimaging experiments. This was achieved with a linear model which used training data to find neural basis images that correspond to the assumed semantic dimensions (for instance, one such basis image might be the activity of the brain for words representing animate concepts), and subsequently used these general patterns and known semantic dimensions to infer the fMRI activity that should be elicited by an unseen stimulus concept. Follow-on work has experimented with other neuroimaging modalities (Murphy et al., 2009), and with a range of semantic models including elicited property norms (Chang et al., 2011), corpus derived models (Devereux and Kelly, 2010; Pereira et al., 2011) and structured ontologies (Jelodar et al., 2010).

The current state-of-the-art performance on this task is achieved using models that are hand-tailored in some respect, whether using manual annotation tasks (Palatucci et al., 2009), use of a domain-appropriate curated corpus (Pereira et al., 2011), or selection of particular collocates to suit the concepts to be described (Mitchell et al., 2008). While these approaches are clearly very successful, it is questionable whether they are a general solution to describe the various parts-of-speech and semantic domains that make up a speaker's vocabulary. The Mitchell et al. (2008) 25-verb model would probably have to be extended to describe the lexicon at large, and it is unclear whether such a compact model could be maintained. While Wikipedia (Pereira et al., 2011) has very broad and increasing cov-

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 114–123,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

erage, it is possible that it will remain inadequate for specialist vocabularies, or for less-studied languages. And while the method used by Palatucci et al. (2009) distributes the annotation task efficiently by crowd-sourcing, it still requires that appropriate questions are compiled by researchers, a task that is both difficult to perform in a systematic way, and which may not generalize to more abstract concepts.

In this paper we examine a representative set of corpus-derived models of meaning, that require no manual intervention, and are applicable to any syntactic and semantic domain. We concentrate on which types of basic corpus pattern perform well on the neurosemantic decoding task: LSA-style **word-region** co-occurrences, and various HAL-style **word-collocate** features including raw tokens, POS tags, and a full dependency parse. Otherwise a common feature extraction and preprocessing pipeline is used: a co-occurrence frequency cutoff, application of a frequency normalization weighting, and dimensionality reduction with SVD.

The following section describes how the brain activity data was gathered and processed; the construction of several corpus-derived models of meaning; and the regression-based methods used to predict one from the other, evaluated with a brain-image matching task (Mitchell et al., 2008). In section 3 we report the results, and in the Conclusion we discuss both the practical implications, and what this works suggests for the cognitive plausibility of distributional models of meaning.

2 Methods

2.1 Brain activity features

The dataset used here is that described in detail in (Mitchell et al., 2008) and released publicly[1] in conjunction with the NAACL 2010 Workshop on Computational Neurolinguistics (Murphy et al., 2010). Functional MRI (fMRI) data was collected from 9 participants while they performed a property generation task. The stimuli were line-drawings, accompanied by their text label, of everyday concrete concepts, with 5 exemplars of each of 12 semantic classes (mammals, body parts, buildings, building parts, clothes, furniture, insects, kitchen utensils, miscellaneous functional artifacts, work tools, vegetables, and vehicles). Stimuli remained on screen for three seconds, and each was each presented six times, in random order, to give a total of 360 image presentations in the session.

The fMRI images were recorded with 3.0T scanner at 1 second intervals, with a spatial resolution of 3x3x6mm. The resulting data was preprocessed with the SPM package (Friston et al., 2007); the blood-oxygen-level response was approximated by taking a boxcar average over a sequence of brain images in each trial; percent signal change was calculated relative to rest periods, and the data from each of the six repetitions of each stimulus were averaged to yield a single brain image for each concept. Finally, a grey-matter anatomical mask was used to select only those voxels (three-dimensional pixels) that overlap with cortex, yielding approximately 20 thousand features per participant.

2.2 Models of semantics

Our objective is to compare current semantic representations that get state-of-the-art performance on the neuro-semantics task with representative distributional models of semantics that can be derived from arbitrary corpora, using varying degrees of linguistic preprocessing. A series of candidate models were selected to represent the variety of ways in which basic textual features can be extracted and represented, including token co-occurrence in a small local window, dependency parses of whole sentences, and document co-occurrence, among others. Other parameters were kept fixed in a way that the literature suggests would be neutral to the various models, and so allow a fair comparison among them (Sahlgren, 2006; Bullinaria and Levy, 2007; Turney and Pantel, 2010).

All textual statistics were gathered from a set of 50m English-language web-page documents consisting of 16 billion words. Where a fixed text window was used, we chose an extent of ±4 lower-case tokens either side of the target

[1] http://www.cs.cmu.edu/afs/cs/project/theo-73/www/science2008/data.html

word of interest, which is in the mid-range of optimal values found by various authors (Lund and Burgess, 1996; Rapp, 2003; Sahlgren, 2006). Positive pointwise-mutual-information (1,2) was used as an association measure to normalize the observed co-occurrence frequency $p(w, f)$ for the varying frequency of the target word $p(w)$ and its features $p(f)$. PPMI up-weights co-occurrences between rare words, yielding positive values for collocations that are more common than would be expected by chance (i.e. if word distributions were independent), and discards negative values that represent patterns of co-occurrences that are *rarer* than one would expect by chance. It has been shown to perform well generally, with both word- and document-level statistics, in raw and dimensionality reduced forms (Bullinaria and Levy, 2007; Turney and Pantel, 2010).[2]

$$\mathrm{PPMI}_{wf} = \begin{cases} \mathrm{PMI}_{wf} & \text{if}\,\mathrm{PMI}_{wf} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$\mathrm{PMI}_{wf} = log\left(\frac{p(w,f)}{p(w)p(f)}\right) \tag{2}$$

A frequency threshold is commonly applied for three reasons: low-frequency co-occurrence counts are more noisy; PMI is positively biased towards hapax co-occurrences; and due to Zipfian distributions a cut-off dramatically reduces the amount of data to be processed. Many authors use a threshold of approximately 50-100 occurrences for word-collocate models (Lund and Burgess, 1996; Lin, 1998; Rapp, 2003). Since Bullinaria and Levy (2007) find improving performance with models using progressively lower cutoffs we explored two cut-offs of 20 and 50 which equate to low co-occurrences thresholds of 0.00125 or 0.003125 per million respectively; for the word-region model we chose a threshold of 2 occurrences of a target term in a document, to keep the input features to a reasonable dimensionality (Bradford, 2008).

After applying these operations to the input data from each model, the resulting dimensionality ranged widely, from about 500 thousand, to tens of millions. A singular value decomposition (SVD) was applied to identify the 1000 dimensions within each model with the greatest explanatory power, which also has the effect of combining similar dimensions (such as synonyms, inflectional variants, topically similar documents) into common components, and discarding more noisy dimensions in the data. Again there is variation in the number of dimension that authors use: here we experiment with 300 and 1000. For decomposition we used a sparse SVD method, the Implicitly Restarted Arnoldi Method (Lehoucq et al., 1998; Jones et al., 2001), which was coherent with the PPMI normalization used, since a zero value represented both negative target-feature associations, and those that were not observed or fell below the frequency cut-off. We also streamlined the task by reducing the input data C (of n target words by m co-occurrence features) to a square matrix CC^T of size $n \times n$, taking advantage of the equality of their left singular vectors U. For SVD to generalize well over the many input features, it is also important to have more training cases that the small set of 60 concrete nouns used in our evaluation task. Consequently we gathered all statistics over a set of the 40,000 most frequent word-forms found in the American National Corpus (Nancy Ide and Keith Suderman, 2006), which should approximate the scale and composition of the vocabulary of a university-educated speaker of English (Nation and Waring, 1997), and over 95% of tokens typically encountered in English.

2.2.1 Hand-tailored benchmarks

The state-of-the-art models on this brain activity prediction task are both hand-tailored. Mitchell et al. (2008) used a model of semantics based on co-occurrence in the Google 1T 5-gram corpus of English (Brants and Franz, 2006) with a small set of **25 Verbs** chosen to represent everyday sensory-motor interaction with concrete objects, such as *see, move, listen.* We recreated this using our current parameters (web document corpus, co-occurrence frequency cut-off, PPMI normalization). The second hand-

[2] Preliminary analyses confirmed that PPMI performed as well or better than alternatives including log-likelihood, TF-IDF, and log-entropy.

tailored dataset we used was a set of **Elicited Properties** inspired by the *20 Questions* game, and gathered using Mechanical Turk (Palatucci et al., 2009; Palatucci, 2011). Multiple informants were asked to answer one or more of 218 questions "related to size, shape, surface properties, and typical usage" such as *Do you see it daily?, Is it wild?, Is it man-made?* with a scalar response ranging from 1 to 5. The resulting responses were then averaged over informants, and then the values of each question were grouped into 5 bins, giving all dimensions similar mean and variance.

2.2.2 Word-Region Model

Latent Semantic Analysis (Deerwester et al., 1990; Landauer and Dumais, 1997), and its probabilistic cousins (Blei et al., 2003; Griffiths et al., 2007), express the meaning of a word as a distribution of co-occurrence across a set of documents, or other text-regions such as paragraphs. This word-region matrix instantiates the assumption that words that share a topical domain (such as medicine, entertainment, philosophy) would be expected to appear in similar sub-sets of text-regions. In such a model, the nearest neighbors of a target word are syntagmatically related (i.e. appear alongside each other), and for *judge* might include *lawyer, court, crime,* or *prison.*

The **Document** model used here is loosely based on LSA, taking the frequency of occurrence of each of our 40,000 vocabulary words in each of 50 million documents as its input data, and it follows Bullinaria and Levy (2007); Turney and Pantel (2010) in using PPMI as a normalization function. We have not investigated variations on the decomposition algorithm in any detail, such as those using non-negative matrix factorization, probabilistic LSA or LDA topic models, as the objective in this paper is to provide a direct comparison between the different types of basic collocation information encoded in corpora, rather than evaluate the best algorithmic means for discovering latent dimensions in those co-occurrences. Nor have we evaluated performance on a more structured corpus input (Pereira et al., 2011). However preliminary tests with the Wikipedia corpus, and with LDA, using the Gensim package (Rehurek and Sojka, 2010) yielded similar performances.

2.2.3 Word-Collocate Models

Word-collocate models make a complementary assumption to that of the document model: that words with closely-related categorical or taxonomic properties should appear in the same position of similar sentences. In a basic word-collocate model, based on a word-word co-occurrence matrix, the nearest neighbors of *judge* might be *athlete, singer,* or *fire-fighter,* reflecting paradigmatic relatedness (i.e. substitutability). Word-collocate models are further differentiated by the amount of linguistic annotation attached to word features, ranging from simple word-form features in a fixed-width window around the concept word, to more elaborate word sequence patterns and parses including parts of speech and dependency relation tags. Among these alternatives, we might expect a dependency model to be the most powerful. Intuitively, the information that *John* is sentient is similarly encoded in the text *John likes cake* and *John seems to really really like cake*, and a suitably effective parser should be able to generalize over this variation, to extract the same dependency relationship of *John-subject-like.* In contrast a narrow window-based model might exclude informative features (such as *like* in the second example), while including presumably uninformative ones (such as *really*). However parsers have the disadvantage of being computationally expensive (meaning that they typically are applied to smaller corpora) and usually introduce some noise through their errors. Consequently, simpler window-based models have often been found to be as effective.

The most basic model considered is the **Word-Form** model, in which all lower-case tokens (word forms and punctuation) found within four positions left and right of the target word are recorded, yielding simple features such as {*john, likes*}. It may also be termed a 'flat' model in contrast to those which assign a variable weight to collocates, progressively lower as one moves further than the target position (e.g.

Lund et al., 1995). We did not use a stop-list, as Bullinaria and Levy (2007) found co-occurrence with very high frequency words also to be informative for semantic tasks. We also expect that the subsequent steps of normalizing with PPMI, reduction with SVD, and use of regularised regression should be able to recognize when such high-frequency words are not informative and then discount these, without the need for such assumptions upfront.

The **Stemmed** model is a slight variation on the Word-Form model, where the same statistics are aggregated after applying Lancaster stemming (Paice, 1990; Loper and Bird, 2002).

The **Directional** model, inspired by Schütze and Pedersen (1993), is also derived from the word-form model, but differentiates between co-occurrence to the left or to the right of the target word, with features such as {*john_L, cake_R*}.

The **Part-of-Speech** model (Kanejiya et al., 2003; Widdows, 2003) replaces each lowercase word-token with its part-of-speech disambiguated form (e.g. *likes_VBZ, cake_NN*). These annotations were extracted from the full dependency parse described below.

The **Sequence** model draws on a range of work that uses word sequence patterns (Lin and Pantel, 2001; Almuhareb and Poesio, 2004; Baroni et al., 2010), and may also be considered an approximation of models that use shallow syntactic analysis (Grefenstette, 1994; Curran and Moens, 2002). All distinct token sequences up to length 4 either side of the target word were counted.

Finally the **Dependency** model uses a full dependency parse, which might be considered the most informed representation of the word-collocate relationships instantiated in corpus sentences, and this approach has been used by several authors (Lin, 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010). The features used are pairs of dependency relation and lexeme corresponding to each edge linked to a target word of interest (e.g. *likes_subj*). The parser used here was Malt, which achieves accuracies of 85% when deriving labelled dependencies on English text (Hall et al., 2007). The features produced by this module are much more limited, to those words that have a direct dependency relation with the word of interest.

2.3 Linear Learning Model

A linear regression model will allow us to evaluate how well a given model of word semantics can be used to predict brain activity. We follow the analysis in Mitchell et al. (2008) and subsequently adopted by several other research groups (see Murphy et al., 2010). For each participant and selected fMRI feature (i.e. each voxel, which records the time-course of neural activity at a fixed location in the brain), we train a model where the level of activation of the latter (the blood oxygenation level) in response to different concepts is approximated by a regularised linear combination of their semantic features:

$$f = \mathbf{C}\beta + \lambda||\beta||^2 \tag{3}$$

where f is the vector of activations of a specific fMRI feature for different concepts, the matrix $\mathbf{C}$ contains the values of the semantic features for the same concepts, β is the vector of weights we must learn for each of those (corpus-derived) features, and λ tunes the degree of regularisation. We can illustrate this with a toy example, containing several stimulus concepts and their attributes on three semantic dimensions: *cat* (*+animate, -big, +moving*); *phone* (*-animate, -big, -moving*); elephant (*+animate, +big, +moving*); skate-board (*-animate, -big, +moving*). After training over all the voxels in our fMRI data with this simple semantic model, we can derive whole brain images that are typical of each of the semantic dimensions. The power of the model is its ability to predict activity for concepts that were not in the training set – for instance the brain activation elicited by the word *car* might be approximated by combining the images see for *-animate, +big, +moving*, even though this combination of properties was not observed during training.

The linear model was estimated with a least squared errors method and *L*2 regularisation, selecting the lambda parameter from the range 0.0001 to 5000 using Generalized Cross-Validation (see Hastie et al., 2011, p.244). The activation of each fMRI voxel in response to a new concept that was not in the training data was predicted by a β-weighted sum of the values on each semantic dimension, building a picture of expected the global neural activity response for an arbitrary concept. Again following Mitchell et al. (2008) we use a leave-2-out paradigm in which a linear model for each neural feature is trained in turn on all concepts minus 2, having selected the 500 most stable voxels in the training set using the same correlational measure across stimulus presentations. For each of the 2 left-out concepts, we predict the global neural activation pattern, as just described. We then try to correctly match the predicted and observed activations, by measuring the cosine distance between the model-generated estimate of fMRI activity and the that observed in the experiment. If the sum of the matched cosine distances is lower than the sum of the mismatched distances, we consider the prediction successful – otherwise as failed. At chance levels, expected matching accuracy is 50%, and significant performance above chance can be estimated using the binomial test, once variance had been verified over independent trials (i.e. where no single stimulus concept is shared between pairs).

3 Results

Table 1 shows the main results of the leave-two-out brain-image matching task. They show the mean classification performance over 1770 word pairs (60 select 2) by 9 participants. All of these classification accuracies are highly significant at $p \ll 0.001$ over test trials (binomial, chance 50%, n=1770*9) and $p < 0.001$ over words (binomial, chance 50%, n=60). There were some significant differences between models when making inferences over trials, but for the small set of words used here it is not possible to make firm conclusions about the superiority of one model over the other, that could be confidently expected to generalize to other stimuli or experiments. However, we do achieve classification accuracies that are as high, or higher than any previously published (Palatucci et al., 2009; Pereira et al., 2011), while models based on very different basic features (directional word-forms; dependency relations; document co-occurrence) yield very similar performance.

Semantic Models	Features	Accuracy
25 Verbs	25	78.5
Elicited Properties	218	83.5
Document (f2)	1000	76.2
Word Form	1000	80.0
Stemmed	1000	76.2
Direction	1000	80.2
Part-of-Speech	1000	80.0
Sequence	1000	78.5
Dependency	1000	**83.1**

Table 1: Brain activity prediction accuracy on leave-2-out pair-matching task. A frequency cutoff of 20 was used for all 1000 dimensional models.

Semantic Models	300 Feats.	1000 Feats.
Document (f2)	79.9	76.2
Word Form	78.1	80.0
Stemmed	77.9	76.2
Direction	80.0	80.2
Part-of-Speech	77.9	80.0
Sequence	72.9	78.5
Dependency	81.6	83.1

Table 2: Effect of SVD dimensionality in the leave-2-out pair-matching setting; frequency cutoff of 20.

3.1 Effect of Number of Dimensions

Here we evaluate what effect the number of SVD dimensions used has on the final performance of various semantic models. Experimental results comparing 300 and 1000 dimensions are presented in Table 2, all based on a frequency cutoff of 20. We observe that performance improves in 5 out of 7 semantic models compared, with the highest performance achieved by the Dependency model when 1000 SVD dimensions were used.

3.2 Effect of Frequency Cutoff

In this section, we evaluate what effect frequency cutoff has on the brain prediction accuracy of various semantic models. From the results in Table 3, we observe only marginal changes as the frequency cutoff varied from 20 to 50. This suggests that the semantic models of this set of

Semantic Models	Cutoff = 50	Cutoff = 20
Document (f2)	79.9	79.9
Word Form	78.5	78.1
Stemmed	78.2	77.9
Direction	80.8	80.0
Part-of-Speech	77.5	77.9
Sequence	74.4	72.9
Dependency	81.3	81.6

Table 3: Effect of frequency cutoff in the leave-2-out pair-matching setting; 300 SVD dimensions.

words are not very sensitive to variations in the frequency cutoff under current experimental settings, and do not benefit clearly from the decrease in sparsity and increase in noise that a lower threshold produces.

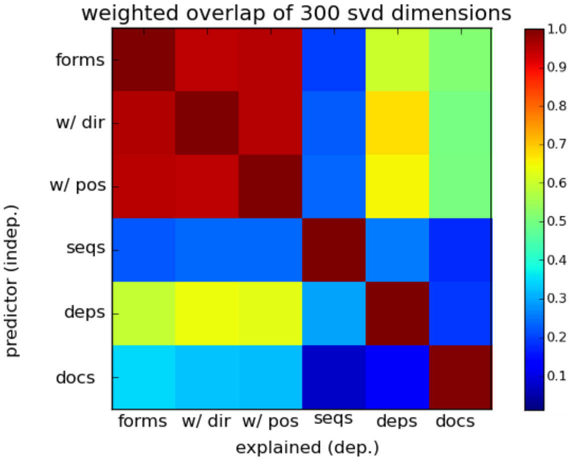
3.3 Information Overlap Analysis

To verify that the models are in fact substantially different, we performed a follow-on analysis that measured the informational overlap between the corpus-derived models. Given two models A and B, both with dimensionality 40 thousand words by 300 SVD dimensions, we can evaluate the extent to which A (used as the predictor semantic representation) contains the information encoded in B (the explained representation). As shown in (4), for each SVD component c, we take the left singular vector b_c as a dependent variable and fit it with a linear model, using the matrix A (all left singular vectors) as independent variables. The explained variance for this column is weighted by its squared singular value s_c^2 in B, and the sum of these component-wise variances gives the total variance explained $R^2_{A\to B}$.

$$R^2_{A\to B} = \sum_{c=1}^{300} \frac{s_c^2}{\sum s_c^2} R_{A\to b_c} \qquad (4)$$

Figure 1 indicates that the first three models, which are all derived from token occurrences in a ±4 window, are close to identical. The sequence and document models are relatively dissimilar, and the dependency model occupies a middle ground, with some similarity to all the models. It is also interesting to note that the among the first cluster of word-form derived models, the directional one has the highest similarity to the dependency model.

Figure 1: Informational Overlap between Corpus-Derived Datasets, in R^2



4 Conclusion

The main result of this study was that we achieved classification accuracies as high as any published, and within a fraction of a percentage point of the human benchmark *20 Questions* data, using completely unsupervised, data-driven models of semantics based on a large random sample of web-text. The most linguistically informed among the models (and so, perhaps the most psychologically plausible), based on dependency parses, is the most successful. Still the performance of sometimes radically different models, from Document-based (syntagmatic) and Word-Form-based (paradigmatic), is surprisingly similar. One reason for this may be that we have reached a ceiling in performance on the fMRI data, due to its inherent noise – in this regard it is interesting to note that an attempt to classify individual concepts using this data directly, without an intervening model of semantics, also achieves about 80% (though on a different task, Shinkareva et al., 2008). Another possible explanation is that both methods reveal equivalent sets of underlying semantic dimensions, but figure 1 suggests not. Alternatively, it may be that the small set of 60 words examined here may be as well-distinguished by means of their taxonomic differences, as by their topical differences, a suggestion supported by the results in Pereira et al. (2011, see Figure 2A).

From the perspective of computational efficiency however, some of the models have clearer advantages. The Dependency and Part-of-Speech models are processing-intensive, since the broad vocabulary considered requires that the very large quantities of text pass through a parsing or tagging pipeline (though these tasks can be parallelized). The Sequence and Document models conversely require very large amounts of memory to store all their features during SVD. In comparison, the Direction model is impressive, as it achieves close to optimal performance, despite being very cheap to produce in terms of processor time and memory footprint. Its relatively superior performance may be due to the relatively fixed word-order of English, making it a good approximation of a Dependency model. For instance, given the narrow ±4 token windows used here, the Direction features *shaky_Left* and *donate_Right* (relative to a target noun) are probably nearly identical to the Dependency features *shaky_Adj* and *donate_Subj*. The Sequence model might also be seen as an approximate Dependency model, but one with the addition of more superficial collocations such as "fish and chips" or "Judge Judy", which are less relevant to our semantic task.

The evidence for the influence of the scaling parameters (number of SVD dimensions, frequency cutoff) is mixed: cut-off appears to have little effect either way, and increasing the number of dimensions can help or hinder (compare the Sequence and Document models). We can speculate that the Document model is already "saturated" with 300 dimensions/topics, but that the other models based on properties have a higher inherent dimensionality. It may also be a lower cut-off and higher dimensionality would show clearer benefits over a larger set of semantic/syntactic domains, including lower-frequency words (the lowest frequency work in the set of 60 used here was *igloo*, which has an incidence of 0.3 per million words in the ANC).

PPMI appears to be both effective, and parsimonious with assumptions one might make about conceptual representations, where it would be cognitively onerous and unnecessary to encode all *negative* features (such as the facts that *dogs* do not have wheels, are not communication events, and do not belong in the aviation domain). But while SVD is certainly effective in dealing with the pervasive synonymy and polysemy seen in corpus-feature sets, it is less clear that it reveals psychologically plausible dimensions of meaning. Alternatives such as non-negative matrix factorization (Lee and Seung, 1999) or Latent Dirichlet Allocation (Blei et al., 2003) might extract more readily interpretable dimensions; or alternative regularisation methods such as Elastic Nets, Lasso (Hastie et al., 2011), or Network Regularisation (Sandler et al., 2009) might even be capable of identifying meaningful clusters of features when learning directly on co-occurrence data. Finally, we should consider whether more derived datasets could be used as input data in place of the basic corpus features used here, such as the full facts learned by the NELL system (Carlson et al., 2010), or crowd-sourced data which can be easily gathered for any word (e.g. association norms, Kiss et al., 1973), though different algorithmic means would be needed to deal with their extreme degree of sparsity.

The results also suggest a series of follow-on analyses. A priority should be to test these models against a wider range of neuroimaging data modalities (e.g. MEG, EEG) and stimulus sets, including abstract kinds (see Murphy et al. 2012, for a preliminary study), and parts-of-speech beyond nouns. It may be that a putative complementarity between word-region and word-collocate models is only revealed when we look at a broader sample of the human lexicon. And beyond establishing what informational content is required to make semantic distinctions, other factorisation methods (e.g. sparse or non-negative decompositions) could be applied to yield more interpretable dimensions. Other classification tasks might also be more sensitive for detecting differences between models, such as the test of word identification among a set by rank accuracy, as used in (Shinkareva et al., 2008).

References

Almuhareb, A. and Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In *Proceedings of EMNLP*, pages 158–165.

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

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–254.

Battig, W. F. and 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 Monographs*, 80(3):1–46.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Bradford, R. B. (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. *Proceeding of the 17th ACM conference on Information and knowledge mining CIKM 08*, pages 153–162.

Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., and Mitchell, T. M. (2010). Toward an Architecture for Never-Ending Language Learning. *Artificial Intelligence*, 2(4):3–3.

Chang, K.-m. K., Mitchell, T., and Just, M. A. (2011). Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fMRI activation. *NeuroImage*, 56(2):716–727.

Curran, J. R. and Moens, M. (2002). Improvements in automatic thesaurus extraction. In *SIGLEX*, pages 59–66.

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

Devereux, B. and Kelly, C. (2010). Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In Murphy, B., Korhonen, A., and Chang, K. K.-M., editors, *1st Workshop on Computational Neurolinguistics*.

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., and Penny, W. D. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, volume 8. Academic Press.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer, Dordrecht.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Hall, J., Nilsson, J., Nivre, J., Eryigit, G., Megyesi, B., Nilsson, M., and Saers, M. (2007). Single Malt or Blended? A Study in Multilingual Parser Optimization. *CoNLL Shared Task Session*, pages 933–939.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning*, volume 18 of *Springer Series in Statistics*. Springer, 5th edition.

Jelodar, A. B., Alizadeh, M., and Khadivi, S. (2010). WordNet Based Features for Predicting Brain Activity associated with meanings of nouns. In Murphy, B., Korhonen, A., and Chang, K. K.-M., editors, *1st Workshop on Computational Neurolinguistics*, pages 18–26.

Jones, E., Oliphant, T., Peterson, P., and Et Al. (2001). SciPy: Open source scientific tools for Python.

Kanejiya, D., Kumar, A., and Prasad, S. (2003). Automatic evaluation of students' answers using syntactically enhanced LSA. *Building educational applications, NAACL*, 2:53–60.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. In Aitken, A. J., Bailey, R. W., and Hamilton-Smith, N., editors, *The Computer and Literary Studies*. Edinburgh University Press.

Landauer, T. and 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.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91.

Lehoucq, R. B., Sorensen, D. C., and Yang, C. (1998). *Arpack users' guide: Solution of large scale eigenvalue problems with implicitly restarted Arnoldi methods*. SIAM.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *COLING-ACL*, pages 768–774.

Lin, D. and Pantel, P. (2001). DIRT – discovery of inference rules from text. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining KDD 01*, datamining:323–328.

Loper, E. and Bird, S. (2002). {NLTK}: The natural language toolkit. In *ACL Workshop*, volume 1, pages 63–70. Association for Computational Linguistics.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical cooccurrence. *Behavior Research Methods, Instruments, and Computers*, 28:203–208.

Lund, K., Burgess, C., and Atchley, R. (1995). Semantic and associative priming in high dimensional semantic space. In *Proceedings of the 17th Cognitive Science Society Meeting*, pages 660–665.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320:1191–1195.

Murphy, B., Baroni, M., and Poesio, M. (2009). EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of EMNLP*, pages 619–627. ACL.

Murphy, B., Korhonen, A., and Chang, K. K.-M., editors (2010). *Proceedings of the 1st Workshop on Computational Neurolinguistics, NAACL-HLT*, Los Angeles. ACL.

Murphy, B., Talukdar, P., and Mitchell, T. (2012). Comparing Abstract and Concrete Conceptual Representations using Neurosemantic Decoding. In *NAACL Workshop on Cognitive Modelling and Computational Linguistics*.

Nancy Ide and Keith Suderman (2006). The American National Corpus First Release. *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*.

Nation, P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N. and McCarthy, M., editors, *Vocabulary Description acquisition and pedagogy*, pages 6–19. Cambridge University Press.

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

Paice, C. D. (1990). Another stemmer. *SIGIR Forum*, 24(3):56–61.

Palatucci, M., Hinton, G., Pomerleau, D., and Mitchell, T. M. (2009). Zero-Shot Learning with Semantic Output Codes. *Advances in Neural Information Processing Systems*, 22:1–9.

Palatucci, M. M. (2011). *Thought Recognition: Predicting and Decoding Brain Activity Using the Zero-Shot Learning Model*. PhD thesis, Carnegie Mellon University.

Pereira, F., Detre, G., and Botvinick, M. (2011). Generating Text from Functional Brain Images. *Frontiers in Human Neuroscience*, 5:1–11.

Rapp, R. (2003). Word Sense Discovery Based on Sense Descriptor Dissimilarity. *Proceedings of the Ninth Machine Translation Summit*, pp:315–322.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *New Challenges, LREC 2010*, pages 45–50. ELRA.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Dissertation, Stockholm University.

Sandler, T., Talukdar, P. P., Ungar, L. H., and Blitzer, J. (2009). Regularized Learning with Networks of Features. *Advances in Neural Information Processing Systems 21*, 4:1401–1408.

Schütze, H. and Pedersen, J. (1993). A Vector Model for syntagmatic and paradigmatic relatedness. In *Making Sense of Words Proceedings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*, pages 104–113.

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., and Just, M. A. (2008). Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. *PloS ONE*, 3(1).

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Artificial Intelligence*, 37(1):141–188.

Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *NAACL*, pages 197–204. Association for Computational Linguistics.