Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora

Barry Devereux
Centre for Speech, Language and the Brain
Department of Experimental Psychology
University of Cambridge
barry@csl.psychol.cam.ac.uk

Colin Kelly & Anna Korhonen
Computer Laboratory
University of Cambridge
{ck329,alk23}@cam.ac.uk

Abstract

We present a series of methods for deriving conceptual representations from corpora and investigate the usefulness of the fMRI data and machine learning methodology of Mitchell et al. (2008) as a basis for evaluating the different models. Within this framework, the quality of a semantic model is quantified by its ability to predict the fMRI activation associated with conceptual stimuli. Mitchell et al. used a manually-acquired set of verbs as the basis for their semantic model; in this paper, we also consider automatically acquired feature-norm-like semantic representations. These models make different assumptions about the kinds of information available in corpora that is relevant to representing conceptual knowledge. Our results indicate that automatically-acquired representations can make equally powerful predictions about the brain activity associated with the stimuli.

1 Introduction

Mitchell et al. (2008) presented a novel approach for predicting human brain activity associated with conceptual stimuli. This approach represents a useful development for interdisciplinary researchers interested in lexical semantics, for several reasons. Most broadly, it is useful in testing the hypothesis that distributional properties of words in corpora can reveal important information about the meanings of words. A strong version of this hypothesis (i.e. that children in part learn the meaning of concrete concept words from co-occurring words in discourse that they are exposed to) has formed the basis of one class of probabilistic cognitive models of conceptual representation (Andrews et al., 2005; Andrews et al., 2009; Steyvers, 2010). Furthermore this approach is useful for testing hypotheses about the kind of co-occurring information that is useful for representing conceptual semantics. In Mitchell et al.'s work (2008), for example, they adopt the position that the meaning of concrete concepts is encoded in the brain with information associated with basic sensory and motor activities (such as actions involving changes to spatial relationships and physical actions performed on objects).

At a more technical level, Mitchell et al.'s fMRI activation data[1] give researchers developing feature-based models of conceptual representation an important benchmark for evaluation. For these researchers, a key problem is the lack of a reasonable "gold standard" against which the quality of the representations generated by a computational model may be evaluated. Previous research has adopted two main approaches to evaluation. Firstly, some models – especially those aiming to extract representations composed of psychologically meaningful semantic feature units, such as Baroni et al. (2009) – have been evaluated against features gathered in large scale property norming studies (e.g. McRae et al. (2005)).[2] By comparing the system output against features elicited by people, this kind of eval-

[1] fMRI data measures changes in oxygen concentrations in the brain. These changes are tied to cognitive processes.

[2] In property norming studies, a group of human subjects are asked to cite features which come to mind for a given concept. These features are compiled by frequency (with a minimum frequency cut-off) to generate a list of features for each concept.

Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics, pages 70–78,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

uation aims to test the psychological validity of computational methods. Furthermore, it allows a fine-grained analysis of performance, for example by revealing the classes of features (part-of, taxonomic, etc) which a given model is particularly good at extracting (Baroni et al., 2008).

However, property norms come with important caveats. One problem is that they tend to over-represent informative or salient information about concepts whilst under-representing other kinds of features. For example, participants report that camels have humps, but not that camels have hearts, even though all participants are likely to have both pieces of information accessible in their representation of the concept CAMEL. If a model is successful in extracting these less salient features, there is no way of evaluating their correctness using property norms. A related issue is that participants can only report verbalizable features, which may not represent the total sum of their conceptual knowledge (Murphy, 2002; McRae et al., 2005).

A second problem with using property norms as the basis of evaluation is that there is often no direct lexical match between feature terms appearing in the system output and the norms. Feature norms are typically normalized such that near-synonymous properties (e.g. *is endangered*, *is an endangered species*, *is almost extinct*, etc., for WHALE) given by different participants are mapped to the same feature label (e.g. *is endangered*). As a consequence, a model may correctly extract *endangered* for WHALE, but other lexical forms of the same feature will not match any feature in the norms. One solution to this is to create an expansion set for each feature which includes its synonyms (Baroni et al., 2008). However, this is only a partial solution because lexical variation in features is not limited to synonyms.

A second approach to evaluating semantic models uses classification or similarity data. For example, Andrews et al. (2009) evaluated their models by calculating cosine similarity scores between semantic representations and using these similarity scores to predict behavioral data which are contingent on the semantic similarity between pairs of concepts (e.g. lexical substitution errors, semantic priming latencies, word-association norms, etc). Although this approach is psychologically motivated, it evaluates a set of extracted features more indirectly than comparison with norm data. In computational linguistics, a similarly indirect evaluation method is to cluster the extracted representations. This approach avoids the difficulties in evaluating individual features; however it only allows consideration along one dimension of the data, namely the similarity between pairs of concepts.

fMRI data such as the Mitchell et al. (2008) dataset offers an advancement over both of these evaluation techniques. Unlike, for example, property norming data, fMRI data offers direct insight into how the brain is functioning in response to given stimuli. Its multidimensional nature makes it easier to inspect what aspects of meaning a particular model is performing strongly or weakly on, and allows for better control of experimental variation. Finally, it avoids the two major issues associated with property norms, which we outlined above.

This paper is structured as follows. In the next section, we briefly describe the models which we used to extract conceptual representations for the 60 concepts in the Mitchell et al. (2008) dataset. In Section 3, we outline our experimental objectives, and the framework we adopt for testing our semantic models. In Section 4, we present the results of our evaluation, which indicate above chance performance for each of the models. Finally, we examine the differences between models by investigating for which concepts prediction of the fMRI activity is poorest, and discuss these differences with respect to the differing assumptions made by the methods.

2 Semantic models

We consider four different semantic models in this paper, which are described briefly below. These models were selected as we were interested in the various kinds of knowledge (part-of-speech, syntactic, and semantic) in corpora available to the extraction process, and the extent to which the use of these types of knowledge can affect the quality of the extracted conceptual representations.

2.1 Mitchell verb-based semantic model

The first semantic model we considered was that of Mitchell et al. (2008). This model assumes that sensory-motor information is an important aspect of conceptual representation, and that the information relevant to a target concept's representation can be estimated from the concept word's frequency of co-occurrence with 25 sensory-motor verbs (*eat*, *manipulate*, *push*, etc) in a very large corpus. Our reimplementation of this method used the co-occurrence statistics provided by Mitchell et al.[3] which were extracted from the Google n-gram corpus consisting of 1 trillion words of web text.

2.2 SVD model

Secondly, we implemented a co-occurrence-based Singular Value Decomposition (SVD) model based on the one described by Baroni and colleagues (Baroni and Lenci, 2008; Baroni et al., 2009). This model combines aspects of both the HAL (Landauer et al., 1998) and LSA (Lund and Burgess, 1996) models in constructing representations for words based on their co-occurrences in texts. A word-by-word co-occurrence matrix was constructed for our corpus, storing how often each target word co-occurred with each context word. The set of context words consisted of the 5,000 most frequent content words (i.e. words not occurring in a stop-list of function words) appearing in the corpus. The set of target words consisted of the 60 concept terms appearing in the fMRI dataset, supplemented with the 10,000 most frequent content words in the corpus (with the exception of the top 10 most frequent words). For calculating co-occurrence frequency between target and context words, the context window was defined by sentence boundaries: two words were considered to co-occur if they appeared in the same sentence[4].

Following Baroni and Lenci (2008), the dimensionality of the target-word × context-word co-occurrence matrix was reduced to 150 columns by singular value decomposition. That is, the singular value decomposition of the co-occurrence matrix was computed and the 150 left singular vectors that accounted for most of the variance, multiplied by the corresponding singular values, were used as the 150-dimensional representation of each target term. Similarity between pairs of target words was calculated as the cosine between their vectors, and for each of the 60 concept words in the experimental stimuli we chose the 200 most similar target words to act as the feature terms extracted by the model. The corpus used with this model was the British National Corpus (BNC) (Leech et al., 1994).

2.3 Novel extraction method

Finally we implemented a novel extraction method, which aims to extract property-norm-like, psychologically meaningful features from corpus data (Kelly et al., 2010). The method aims to extract semantically unconstrained feature triples of the form *concept-relation-feature* , *w*here *feature* is a feature (either noun or adjective) of the target concept and *relation* is a verb representing the semantic relationship between them. Examples of extracted triples include: *swan be white*, *swan have neck* and *screwdriver be tool*. The model uses a corpus parsed for grammatical relations (GRs) using Robust Accurate Statistical Parsing (RASP) (Briscoe et al., 2006). For each sentence containing a target concept, the set of GRs for that sentence are examined to test whether they match manually-created rules. These rules include prototypical feature-relation GR structures connecting elements of the sentence and represent dependency patterns which encode potential semantic relationships between the concept and candidate feature terms occurring in the sentence. A large set of candidate triples are extracted by applying these rules to each sentence in the corpus containing a target concept, and the triples for each concept are ranked by their frequency of extraction. In the second stage of the method, the extracted triples are reweighted on the basis of probabilistic high-level semantic information obtained from human property norm data. This subsequent stage has the effect of increasing the weight associated with more high-quality features and downgrading lower-quality features. The extraction method is described more fully in Kelly et al. (2010). For this method we also used the BNC. The top 200 triples ranked by frequency (i.e. unweighted) and the top 200 features after reweighting with the semantic data were used in our experiments.

[3] `http://www.cs.cmu.edu/~tom/science2008/semanticFeatureVectors.html`

[4] In Baroni et al.'s implementation a context window of 5 (Baroni and Lenci, 2008) or 20 (Baroni et al., 2009) words either side of the target word was used instead; we chose a sentence-based context window as it is analogous to the context used in our experimental method (described in the following section).

3 Experiment

As mentioned above, we are primarily interested in using the fMRI data to evaluate the quality of the different methods for extracting conceptual representations from corpora (rather than being interested in investigating methods for predicting fMRI activation). We make no attempt to build on the method described by Mitchell et al. (2008), although there are likely to be many interesting avenues through which that method could be extended.[5] We therefore followed the Mitchell et al. methodology as closely as possible, using the same multiple regression training and leave-two-out cross-validation paradigms as presented in their paper and supporting online material. The only parameter that we varied was the extraction method (and corpus) that was used to generate the feature-vectors associated with the 60 concepts that were used during the training phase. The quality of the predictions generated for the concepts using each semantic model can therefore be adopted as an index of model performance.

The Mitchell et al. method uses co-occurrence with a specific set of 25 manually selected verbs (*eat*, *push*, etc) that are the same for each concept. This results in 25-dimensional feature vectors for input into training. However, for both the SVD model and our triple extraction models there are no *a priori* constraints on the number of unique features that can be extracted for the concepts. For these models, we selected the top 200 features associated with each concept; therefore, across all 60 concepts in the Mitchell et al. dataset, there are thousands of unique features extracted which are used in the concepts' representations. To ensure that the linear regression model for each method would be fitted using the same number of free parameters during training (thereby maximizing the comparability of the different methods), we reduced the dimensionality of the generated feature spaces for the SVD method and the two triple-extraction methods using Principal Components Analysis (PCA). The concept × feature extraction frequency matrices for the three models were submitted to PCA, and the first 25 components (i.e. those components which best characterized the variance of the original features) for each model were selected. In the case of the SVD model, these 25 dimensions explained 77.7% of the variance in the original 3,061-dimensional vectors. For our unweighted extraction method, the 25 extracted components explained 63.0% of the original 5,525 dimensions; for the weighted method the components explained 71.5% of the original 6,567 dimensions.

Triples (weighted)		***SVD***	
PCA1	**PCA2**	**PCA1**	**PCA2**
	Highest-valued concepts		
horse	house	coat	butterfly
cat	apartment	skirt	cow
cow	dog	shirt	ant
dog	igloo	pants	bee
beetle	car	dress	lettuce
	Lowest-valued concepts		
knife	pants	car	desk
door	coat	watch	arm
hammer	dress	horse	chair
saw	skirt	dog	knife
chisel	shirt	fly	leg

Table 1: Highest- and lowest-valued concepts for the first two components for the SVD and weighted triple-extraction methods.

It is interesting to consider the kind of semantic information that is being captured by the resultant PCA components. In particular, the components appear to capture meaningful distinctions between stimuli. For example, the first PCA component for our weighted triple extraction method can be interpreted as the concepts' degree of "animalness" (animal stimuli have high values on this component). Table 1 presents the five highest and lowest-valued concepts for the first two components for the SVD model and the weighted triple extraction model. Concepts which overlap with respect to a specific set of semantic properties tend to have high or low values on a given dimension, indicating that that component is capturing a specific cluster of co-occurring semantic features. For example, PCA1 for SVD can be interpreted as "has features associated with clothing".

Therefore, a key difference between the Michell

[5] For example, the method currently makes the simplifying assumption that the activity in neighbouring voxels is independent.

Method	Feature Type	POS	Syntax	Semantics
Mitchell	25 verbs	no	no	no
SVD	tuples (content-words)	yes	no	no
triple-extraction method (unweighted)	feature-triples	yes	yes	no
triple-extraction method (weighted)	feature-triples	yes	yes	yes

Table 2: Comparison of the information available to each model.

et al. model and our models is that while Mitchell et al. posit that certain sensory-motor function verbs can act as important features of concepts, our models instead place more importance on intrinsic semantic features.

Finally, Table 2 gives a summary comparison of the different models, in terms of whether or not each uses part of speech (POS) data, syntactic information (i.e. GRs), and semantic filtering (Section 2.3).

It should be noted that the BNC corpus (used with the SVD model and our triple-extraction method) is 10,000 times smaller than the corpus from which the Mitchell et al. feature vectors are derived. As such the semantic representations we extract with our method need to make better use of the data available in the corpus if they are to compete with the verb-based features used by Mitchell et al.'s method.

4 Results

The accuracy for each of the four methods was evaluated using a leave-two-out validation paradigm. There are 1,770 possible pairs of concepts that can be drawn from the set of 60 concept stimuli. Training was performed separately for each participant and for each of the 1,770 held-out pairs. Given a particular participant and held-out pair, for each voxel v we fit the activation at that voxel to the set of 58 training items with multiple linear regression, using as predictor variables the elements of the 25-dimensional feature vectors associated with each of the 58 concepts. Training therefore yields a set of 25 β-coefficients, which can be used to generate a prediction for the activation y_v of voxel v for the held-out word w using the equation

$$y_v^{\text{pred}} = \sum_{i=1}^{25} \beta_{v,i} f_{i,w} \qquad (1)$$

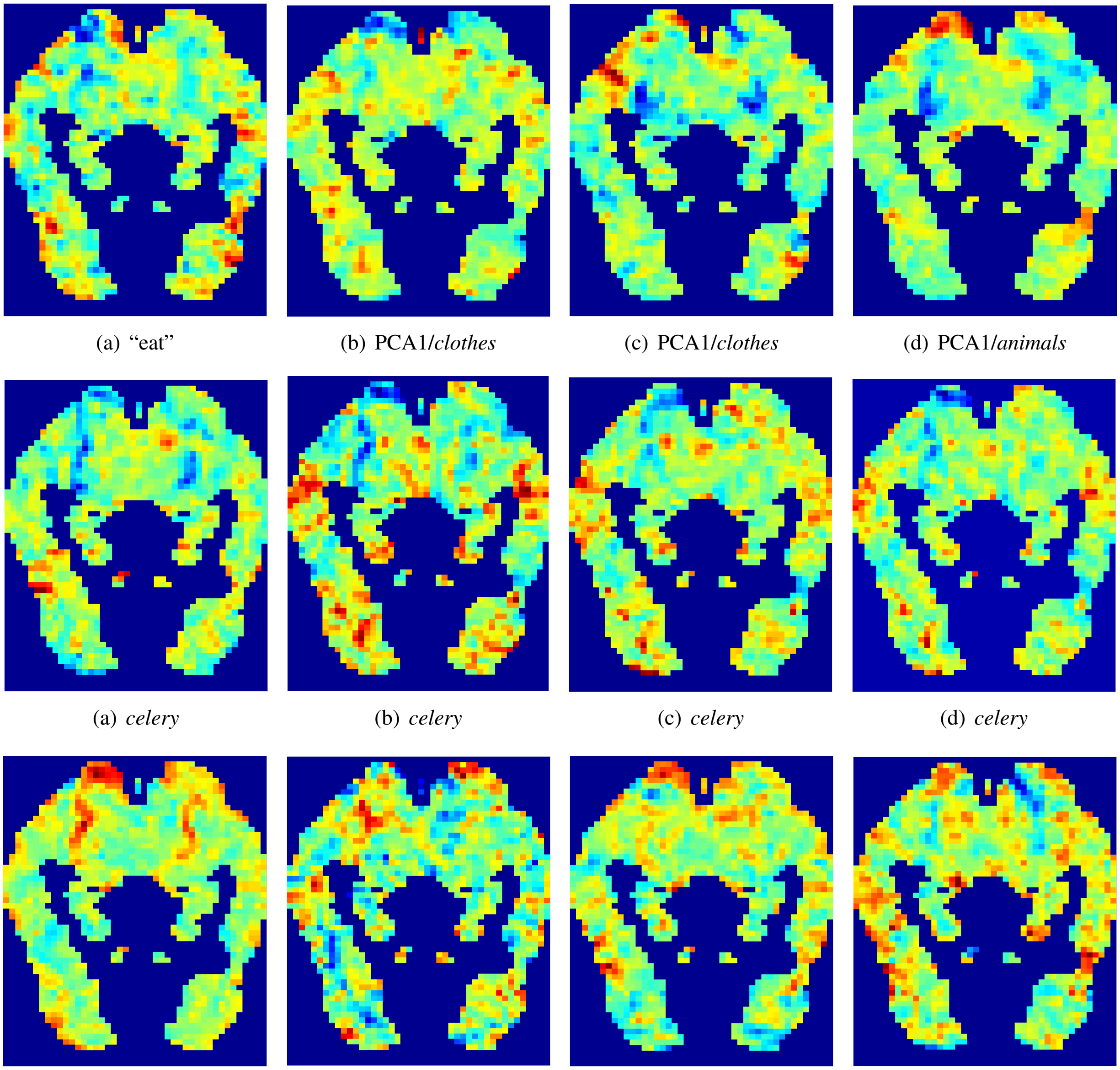
where $f_{i,w}$ is the i^{th} element of the feature vector for word w (see Mitchell et al. (2008) for details). Over all voxels, this method gives a prediction for the activation with respect to the held-out word w which can then be compared to the observed activation for that stimulus.

Rather than comparing the activity between predicted and observed images using all voxels, we compared images using only the 500 most stable voxels for each participant. For each participant, the 500 most stable voxels were the voxels which gave the most consistent pattern of activation across the six presentations of all 60 stimuli (see Mitchell et al. (2008) for details).

The top row of Figure 1 presents the learned coefficients for one feature dimension for each of the four semantic models considered in our experiments (for these images, all voxels rather then the 500 most stable voxels are used). For the Mitchell et al. method, the coefficients presented correspond to the verb *eat*; for the other models the feature is the PCA component that explained the most variance in the original representations. We also present the predicted images for the concepts CELERY and AIRPLANE, calculated on the coefficients learned over the remaining 58 concepts. Importantly, for the Mitchell et al. method (column (a)), the learned coefficients for *eat* and the predicted images for CELERY and AIRPLANE agree with those reported by Mitchell et al. (2008, Figure 2 & online supplementary material[6]).

We calculated similarity between predicted and observed images using both cosine and Pearson correlation and the 500 most stable voxels; we report the results using Pearson correlation here as this measure consistently gave slightly better accuracies

[6] `http://www.cs.cmu.edu/~tom/science2008/featureSignaturesP1.html`



(a) “eat” (b) PCA1/*clothes* (c) PCA1/*clothes* (d) PCA1/*animals*

(a) *celery* (b) *celery* (c) *celery* (d) *celery*

(a) *airplane* (b) *airplane* (c) *airplane* (d) *airplane*

Figure 1: Learned coefficients on a selected feature dimension (top row) and predicted activation for CELERY (middle row) and AIRPLANE (bottom row) for four semantic models: (a) Mitchell et al. (2008), (b) SVD (c) triple extraction method (unweighted), and (d) triple extraction method (weighted). Warmer colours indicate higher values (i.e. larger β-coefficients for the feature dimensions and higher predicted activation for the concepts). PCA components have been given intuitive labels indicating the kind of information described by that component (see Table 1). As in Figure 2 of Mitchell et al. (2008), the figure shows just one slice in the horizontal plane (z = -12 in MNI space) for one participant (P1). The predicted images for CELERY and AIRPLANE were generated from the feature coefficients learned on the other 58 concepts using each of the four models; the corresponding observed images for CELERY and AIRPLANE can be found in Mitchell et al. (2008) Figure 2 B.

Method	P1	P2	P3	P4	P5	P6	P7	P8	P9	Mean
Mitchell et al. (2008)	0.84	0.83	0.76	0.81	0.79	0.66	0.73	0.64	0.68	0.75
SVD	0.82	0.67	0.79	0.83	0.74	0.64	0.64	0.70	0.75	0.73
Triple-extraction (unweighted)	0.82	0.71	0.79	0.80	0.70	0.69	0.65	0.53	0.78	0.72
Triple-extraction (weighted)	0.82	0.72	0.76	0.83	0.73	0.65	0.68	0.51	0.76	0.72

Table 3: Accuracy results for the four semantic models.

for each of the four models (the results are very similar using the cosine measure). Following Mitchell et al. (2008; supplementary material), a match score for each held out pair w_1 and w_2 was calculated as the sum of the similarities between the correctly aligned predicted and observed images:

$$a = sim(w_1^{\text{pred}}, w_1^{\text{obs}}) + sim(w_2^{\text{pred}}, w_2^{\text{obs}}) \qquad (2)$$

Similarly a mismatch score was calculated as

$$b = sim(w_1^{\text{pred}}, w_2^{\text{obs}}) + sim(w_2^{\text{pred}}, w_1^{\text{obs}}) \qquad (3)$$

Cases where the match score is greater than the mismatch score (i.e. $a > b$) count as successes for the model (i.e. the model correctly identifies the two predicted images). Otherwise there is a failure by the model (i.e. the model identifies the observed image for w_1 as being w_2 and vice-versa).

Table 3 presents the results of the leave-two-out cross-validation evaluation, giving the proportion (across all 1,770 pairs) of predicted images for the held-out pairs that were correctly matched to the observed images.[7] The original Mitchell et al. (2008) model has the best mean performance, although across the nine participants, there is no significant difference in accuracy between any of the models ($|t(8)| < 1.49$, $p > 0.17$, for all pairwise paired t-tests between Mitchell et al. (2008), SVD, and weighted triple extraction).

That there is no difference between the performance of the Mitchell et al. (2008), SVD and triple extraction methods is surprising, given the different kinds of information that are available to the different models. In particular, the models that automatically acquire very general and semantically unconstrained feature-based representations perform as well as the model which uses a set of manually-selected sensory-motor verbs, even though the representations generated for these models are derived from 10,000 times less corpus data.

As mentioned in our introduction, an advantage of evaluating against the fMRI dataset is that this multi-dimensional data allows us to investigate strengths and weaknesses of different models in a way which is not possible using similarity or clustering-based evaluation. As a very simple investigation of specific differences in model performance, we present in Table 4 the pairs of concepts for which each of the models performs most poorly on. The Mitchell et al. (2008) method appears to do poorly on pairs of concepts where a constituent word can be ambiguous with respect to its part-of-speech (e.g. SAW, BEAR). This is not surprising, given that part-of-speech data is not available in the Google n-gram corpus used with this method. The performance of the Mitchell et al. method might therefore be improved significantly by applying heuristics to the n-gram data to make inferences about the correct part-of-speech of instances of words like SAW and BEAR. For the SVD and weighted triple extraction methods, which both use the BNC corpus, there is some evidence that the models are performing poorly for relatively low frequency words[8] (e.g. CHISEL), words which are semantically ambiguous as nouns (e.g. ARM), and pairs which are semantically similar (e.g. SPOON & KNIFE). This suggests that the SVD and triple extraction methods may perform better with a larger and more diverse corpus.

[7] Our results for the Mitchell et al. (2008) method are similar, though not identical, to those reported in that paper (where the reported mean accuracy across all participants is 0.77, using cosine similarity). Our implementation of the method for selecting the 500 most stable voxels yields slightly different voxels from those obtained by Mitchell et al. (2008; see supplementary material). In any case, the same set of 500 voxels for each participant were used for generating the results of each model presented here, and so we do not believe that this discrepancy affects comparison of the different models.

[8] AIRPLANE is relatively low frequency in the BNC; it may be more sensible to use the word AEROPLANE with a British corpus.

Mitchell et al.		SVD		Triple Extraction (weighted)	
Pair	*Nr.*	*Pair*	*Nr.*	*Pair*	*Nr.*
bear saw	0	cup airplane	0	dresser chimney	0
bell carrot	0	cup lettuce	0	airplane chisel	0
bell saw	0	horse beetle	0	airplane hand	0
knife bear	0	chisel arm	0	airplane tomato	0
cup saw	1	hammer arm	1	spoon chisel	0
bear tomato	1	dresser arch	1	spoon knife	0

Table 4: Leave-out pairs for which each model performs least accurately, across the nine participants. *Nr.* = the number of participants for which this leave-out pair was correctly matched.

5 Conclusion

The fMRI dataset and training and evaluation methodology presented by Mitchell et al. (2008) gives researchers an interesting new framework with which to evaluate the quality of feature-based conceptual representations extracted from corpora. This framework avoids some of the problems inherent in evaluating extracted representations against a "gold standard" based on participant-generated property norms. It also provides a rich multi-dimensional dataset through which the strengths and weaknesses of extraction methods can be identified.

We have applied this evaluation framework to four feature extraction methods which use different sources of information available in corpora to extract conceptual representations. Surprisingly, in spite of their major differences, we did not find any significant difference in performance between the models.

This finding has interesting theoretical implications, given that previous research has suggested that aspects of meaning defined by sensory-motor verbs may have a somewhat distinctive role to play in predicting the fMRI activation associated with conceptual stimuli (Mitchell et al., 2008). Our results suggest that general feature-based representations of concepts, which place no *a priori* distinction on sensory-motor properties, may be equally capable of predicting activation to conceptual stimuli. This highlights the potential for the Mitchell et al. method to be used to inform both distributed and sensory-motor accounts of conceptual representation (e.g. McRae et al. (1997), Cree et al. (2006), Tyler et al. (2000), Tyler & Moss (2001), Moss et al. (2007), Martin & Chao (2001)), as well as providing a benchmark with which to assess semantic model development. In a similar vein, Murphy et al. (2009) used a dependency-parsed corpus yielding verb co-occurrence statistics to predict EEG[9] activation patterns with significant accuracy.

The training and evaluation framework presented by Mitchell et al. (2008) represents just one point in a large space of possibilities for using computational modelling to predict human brain activity associated with conceptual stimuli. In these initial experiments, we have chosen to follow the Mitchell et al. approach as closely as possible, in order to maximize comparability with their results. In future work, we aim to investigate other methods for training and evaluation, other corpora and other sources of imaging data. Furthermore, we aim to use the evaluation results from such work to inform the development of our extraction method.

Acknowledgments

Our work was funded by the EPSRC grant EP/F030061/1, and the Royal Society University Research Fellowship, UK. We thank Mitchell et al. (2008) and McRae et al. (2005) for making their data publically available.

References

Mark Andrews, G. Vigliocco, and D. Vinson. 2005. Integrating attributional and distributional information in a probabilistic model of meaning representation. In Timo Honkela et al., editor, *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Rea-*

[9] EEG measures voltages induced by neuronal firing across the human scalp.

soning, pages 15–25, Espoo, Finland: Helsinki University of Technology.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Marco Baroni and Alessandro Lenci. 2008. Concepts and properties in word spaces. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics)*, 20(1):55–88.

Marco Baroni, Stefan Evert, and Alessandro Lenci, editors. 2008. *ESSLLI 2008 Workshop on Distributional Lexical Semantics*.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2009. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, pages 1–33.

E. Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the Interactive Demo Session of COLING/ACL-06*, pages 77–80.

George S. Cree, Chris McNorgan, and Ken McRae. 2006. Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 32(4):643–58.

Colin Kelly, Barry Devereux, and Anna Korhonen. 2010. Acquiring human-like feature-based conceptual representations from corpora. In Brian Murphy, Kai min Kevin Chang, and Anna Korhonen, editors, *Proceedings of the NAACL-HLT Workshop on Computational Neurolinguistics*, Los Angeles, USA.

T.K. Landauer, P.W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

G. Leech, R. Garside, and M. Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Alex Martin and Linda L. Chao. 2001. Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, 11(2):194–201.

Ken McRae, Virginia R. de Sa, and Mark S. Seidenberg. 1997. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99–130.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel A. Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Helen E. Moss, Lorraine K. Tyler, and Kirsten I. Taylor. 2007. Conceptual structure. In M. Gareth Gaskell, editor, *The Oxford handbook of psycholinguistics*, pages 217–234. Oxford University Press, Oxford, UK.

B. Murphy, M. Baroni, and M. Poesio. 2009. Eeg responds to conceptual stimuli and corpus semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 619–627, East Stroudsburg, PA.

Gregory Murphy. 2002. *The big book of concepts*. The MIT Press, Cambridge, MA.

Mark Steyvers. 2010. Combining feature norms and text data with topic models. *Acta Psychologica*, 133(3):234–243.

Lorraine K. Tyler and Helen E. Moss. 2001. Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, 5(6):244–252.

L. K. Tyler, H. E. Moss, M. R. Durrant-Peatfield, and J. P. Levy. 2000. Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75(2):195–231.