Experiential, Distributional and Dependency-based Word Embeddings have Complementary Roles in Decoding Brain Activity

Samira Abnar **Rasyan Ahmed** **Max Mijnheer** **Willem Zuidema**
University of Amsterdam
{samiraabnar,rasyan21,max.mijnheer}@gmail.com, zuidema@uva.nl

Abstract

We evaluate 8 different word embedding models on their usefulness for predicting the neural activation patterns associated with concrete nouns. The models we consider include an experiential model, based on crowd-sourced association data, several popular neural and distributional models, and a model that reflects the syntactic context of words (based on dependency parses). Our goal is to assess the cognitive plausibility of these various embedding models, and understand how we can further improve our methods for interpreting brain imaging data.

We show that neural word embedding models exhibit superior performance on the tasks we consider, beating experiential word representation model. The syntactically informed model gives the overall best performance when predicting brain activation patterns from word embeddings; whereas the GloVe distributional method gives the overall best performance when predicting in the reverse direction (words vectors from brain images). Interestingly, however, the error patterns of these different models are markedly different. This may support the idea that the brain uses different systems for processing different kinds of words. Moreover, we suggest that taking the relative strengths of different embedding models into account will lead to better models of the brain activity associated with words.

1 Introduction

How are word meanings represented in the human brain? Is there a single amodal semantic system or are there multiple responsible for representing meanings of different classes of words? Recently, a series of studies have emerged showing that a combination of methods from machine learning, computational linguistics and cognitive neuroscience are useful for addressing such questions.

(Mitchell et al., 2008) pioneered the use of corpus-derived word representations to predict patterns of neural activation's when subjects are exposed to a stimulus word. Using their framework, a series of papers have evaluated various techniques of computing word representation models based on different assumptions, as we review in section 2.

Since these early successes, a range of new word embedding methods have been proposed and successfully used in a variety of NLP tasks, including methods based on deep learning with neural networks. (Baroni et al., 2014) and (Pereira et al., 2016) present systematic studies, showing that also behavioural data from psycholinguistics can be modelled effectively using neural word embedding models such as GloVe(Pennington et al., 2014) and word2vec(Mikolov et al., 2013). At the same time, studies in the area of vision have shown that deep learning models fit very well to the neocortical data (Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014) and they can help to better understand the sensory cortical system (Yamins and DiCarlo, 2016). To investigate how well the new word embedding models, and in particular the deep learning models, fare in helping to understand neural activation patterns in the domain of language, we now present a systematic evaluation of 8 word embedding models, listed in section 3, against the neuroimaging data from (Mitchell et al., 2008), following the experiments and primary results in (Mijnheer, 2017; Ahmed, 2017).

To address this goal, we take word embedding models designed based on different assumptions of how meanings of words can be represented and evaluate their performance on either the task of predicting brain data from word embeddings or the reverse, predicting word embeddings from brain data. The basic assumption here is that the better the performance of a model

Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018), pages 57–66,
Salt Lake City, Utah, USA, January 7, 2018.

is the more probable it is that the way the word embedding model is built reflects what happens in the human brain to understand a meaning of a word. In our experiments, we compare modern neural word embedding models with traditional approaches that are based on manually assigned linguistic word attributes, and neuro-inspired techniques based on sensory-motor features. Besides a large-scale evaluation of various word embedding models, we conduct a detailed error analysis to understand the differences between them.

The first research question we investigate is: *How well does each word embedding model allow us to predict neural activation patterns in human brain?* To answer this we measure how well different word embedding models can predict the brain imaging data. Taking this one step further, we also train our models in the reverse direction: to directly predict word embeddings from brain data.

The second research question that we investigate is: *What is the best word embedding model for predicting brain activation for different (classes of) nouns?* Maybe human brain uses different processes to understand meanings of different kind of words (Riddoch et al., 1988; Caramazza et al., 1990; Warrington and Shallice, 1984; Caramazza and Shelton, 1998). We do a qualitative analysis of our results to see whether different word embedding models are good in predicting the brain activation for different categories of nouns. The third question we address is *Which are the most predictable voxels in the brain for each word embedding model?* By answering this question we want to test the hypothesis that different areas of the brain are responsible for processing different aspect of the meaning of nouns. If different models have different performance either for different noun pairs or for different brain areas, the next step would be to find a way to integrate different models to build a model that better fits the brain data.

2 Related Work

The tradition of developing computational models to predict neural activation patterns given a representation of a stimulus such as a word was started by (Mitchell et al., 2008), who presented a model that quite successfully (with performance well above chance) predicted neural activation patterns associated with nouns, using a hand-designed set of 25 verbs (reflecting sensory-motor features) and computing representations for the nouns based on their co-occurrences with these verbs in a trillion-token corpus. Following this work, (Jelodar et al., 2010) proposed using WordNet (Miller, 1995) instead of corpus statistics to compute the values for the 25 features introduced in (Mitchell et al., 2008), allowing them to deal with some of the ambiguity related issues. They find that a linear combination of their WordNet-based 25 features and the co-occurrence based 25 features of (Mitchell et al., 2008) improves the fMRI neural activity prediction accuracy. Devereux et al (Devereux et al., 2010) applied the framework to evaluate four different feature extraction methods, each based on a different source of information available in corpora. They show that general computational word representation models can be as good as sensory-motor based word representations. Later Murphy et al have done an extensive study comparing the performance of a different kind of corpus-based models on this task. In their experiments, a model that exploits dependency information outperforms the others (Murphy et al., 2012), in line with the results that we report below. (Binder et al., 2016) argue that it makes more sense to use experiment based word representations to model the mental lexicon. In (Fernandino et al., 2015) they use sensory-motor experience based attributes as elements of the word vectors to predict neural activation pattern for lexical concepts. The main difference of this approach with (Mitchell et al., 2008) is that rather than statistics from corpora they use actual human ratings to compute the feature values.

More recently, the success of neural network based approaches for learning word representations has raised the question whether these models might be able to partly simulate how our brain is processing language. Hence, it is now the time to revisit the challenge Tom Mitchell introduced and evaluate these new models with human brain neural activation patterns. In (Anderson et al., 2017) the performance of word2vec as the word representation model for predicting brain activation patterns is already evaluated. The goal of their experiment was to compare a text-based word representation with image-based models; our goal, instead, is to compare different neural word embedding models that are all text-based. Furthermore, (Xu et al., 2016) they compare the performance of various word embedding models, including neural based models and non-distributional models for both behavioural tasks and brain image datasets.

Taking the differences between all these different models for word representation into account, one can argue that they are not replaceable with each other. In (Dove, 2009) it is argued that both perceptual and non-perceptual features are important in decoding

semantics. Moreover (Andrews et al., 2009) has suggested combining experiential and distributional models to learn word representations. In our experiments, we want to investigate whether the information encoded in different kind of word representations are mutually exclusive and hence, integrating them would result in a more powerful model.

There have also been some efforts to extend these models to analyze and understand brain activation patterns at sentence level (Wehbe et al., 2014) or at least in the context of a sentence rather than an isolated word (Anderson et al., 2016a). Moreover, some other related work abstracts away from the brain activation patterns and instead analyzes the correlation between the pairwise similarity of word representations in the brain and the computational model under evaluation (Anderson et al., 2016b).

In this paper, we stay with the original setup, using word representation models for predicting fMRI neural activation patterns, but go beyond existing work by presenting a systematic analysis and comparison of the performance of different kind of word representation models.

3 Experimental Setup

The main task in our experiments is to use a regression model to map word representations to brain activation patterns or vice versa. As the regression model, we employ a single layer neural network with $tanh$ activation. To avoid over-fitting we use drop-connect (Wan et al., 2013) with a keeping rate of 0.7 beside L2 regularization with $\lambda = 0.001$. In all the experiments we train the models for each subject separately. The training and evaluation are done with the leave-2-out method as suggested in (Mitchell et al., 2008). Where we train the model on all except 2 pairs and then evaluate the performance of the model on the left-out pairs. We do this for all possible combinations of pairs.

Neuroimaging Data Our experiments are conducted on the data from Mitchell et al. (2008) which is publicly available[1]. This is a collection of fMRI data that is gathered from 9 participants while exposed to distinctive stimuli. The stimuli consisted of 60 nouns and corresponding line drawings. Each stimulus was displayed six times for 3 seconds in random order, adding to a total of 360 fMRI images per participant.

Word Embedding Models In order to get insights about how human mental lexicon is built, we use a wide variety of recently proposed word representation models. The word embedding models that we are exploring in our experiments are in two (non-exclusive) categories: experiential or distributional. In the experiential model, the meanings of the words are coded to reflect how the corresponding concept is experienced by humans through their senses. In the distributional models, the meaning of words is represented based on their co-occurrence with other words. These models can be either count-based or predictive (Baroni et al., 2014). The word representation models we will use are:

- **Experiential word representations**: Experiential word representations are suggested based on the fact that humans remember the meaning of things as they experience them. In (Binder et al., 2016) a set of 65 features are defined and crowdsourcing is used to rate the relatedness of each feature for each word. Thus, instead of computing the value of features using statistical data from textual corpora they use actual human ratings. We use the dataset introduce in (Binder et al., 2016). Since it contains only about 50% of the nouns in Tom Mitchell et al dataset, some of the experiments we report are with this limited noun set.

- **Distributional word embedding models**:
  - **Word2Vec**: Word2vec basically is a shallow, two layer, neural network that reconstructs the context of a given word. In our experiments, we use the skip gram word2vec model trained on Wikipedia (Mikolov et al., 2013).
  - **Fasttext**: Fasttext is a modification of word2vec that takes morphological information into account (Bojanowski et al., 2016).
  - **Dependency-based word2vec**: The dependency-based word2vec introduced in (Levy and Goldberg, 2014) is a word2vec model in which the context of the words is computed based on the dependency relations.
  - **GloVe**: GloVe is a count-based method. It does a dimensionality reduction on the co-occurrence matrix(Pennington et al., 2014).
  - **LexVec**: LexVec is also a count based method. It is a matrix factorization method that combines ideas from different models. It minimizes the reconstruction loss function that weights frequent co-occurrences heavily while taking into account negative co-occurrence (Salle et al., 2016b,a).

- **25 verb features**: Similar to experiential word

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

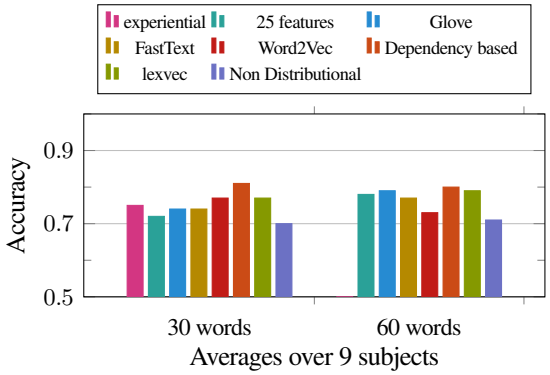


Figure 1: Results for the word to brain activation prediction task. (Chance is .5)

representations, this model is based on the idea that the neural representation of nouns is grounded in sensory-motor features. They have manually picked 25 verbs and suggested to use the co-occurrence counts of nouns with these 25 verbs to form the word representations (Mitchell et al., 2008).

- **non-distributional word vector representation**: (Faruqui and Dyer, 2015) have constructed a non-distributional word representation model employing linguistic resources such as WordNet(Miller, 1995), FrameNet(Baker et al., 1998) etc. In this model, words are presented as binary vectors where each element of the vector indicates whether the represented word has or does not have a specific feature. As a result, the vectors are highly sparse. The advantage of this model to distributional word representations is the interpretability of its dimensions.

4 *How well does each word embedding model allow us to predict neural activation patterns in human brain?*

To address the first research question, we train a separate regression model for each word representation model to compute the average brain activation corresponding to each word for a particular subject. Figure 1 illustrates the results of evaluating these models on the brain activation prediction task, using the leave-2-out methodology we discussed in section 3. For the sake of including the experiential word representations from (Binder et al., 2016) in our evaluations, we also conducted a set of experiments with only the nouns that were included in the experiential word representation collection. The good news is that all the models we are evaluating perform significantly above chance. The fact that the ranking of the models differs per subject makes it difficult to make general conclusions about the best model. Overall, dependency-based word2vec, GloVe and 25 features model are the top-ranked models for at least one of the subjects.

Among neural word embedding models, dependency-based word2vec is achieving the best accuracy. This is in line with the results from (Murphy et al., 2012), where they showed that the corpus-based model considering the dependency relationships has the highest performance among corpus-based models. These authors report an accuracy of 83.1 (with 1000 dimensional word vectors). Somewhat higher still than the best dependency based word2vec, and the highest performance reported in the literature until now for a corpus-based model. The fact that fasttext and dependency based word2vec are performing better than word2vec might reflect the importance of morphological and dependency information. Comparing predictive models with count-based models, although count-based methods like GloVe and LexVec are beating simple word2vec, looking at the performances of fasttext and dependency based word2vec, we can conclude that the context prediction models can potentially perform better. Moreover, comparing the performance of the Experiential Model with 25 feature model, we see that the Experiential Model is doing slightly better on average while their ranking is different per subject. Either the higher number of features or the way feature values are computed could have led to the slight improvement in accuracy for the experiential model.

In both sets of experiments in Figure 1 the non-distributional word representation model has the lowest performance. The very high dimensionality of the brain imaging data versus the sparseness of non-distributional word vectors make training the regression model with these vectors much harder and this might be the primary reason for its low performance.

Next, instead of predicting brain activation patterns, we train the regression model to predict the word representation given a brain activation. Thus, we want to predict the stimulus word from the neural activation pattern in the brain. Evaluation is still based on the leave-2-out setup (so we still evaluate with 2 brain images and 2 word embeddings at each instance, making quantitative results comparable across experiments).

The results are shown in Figure 3. We expected the performance of the models on the reversed task, predicting word features from brain activation, to be somewhat similar to their performance on the main task, predicting brain activation patterns from word vectors. However, the results are surprising. For the

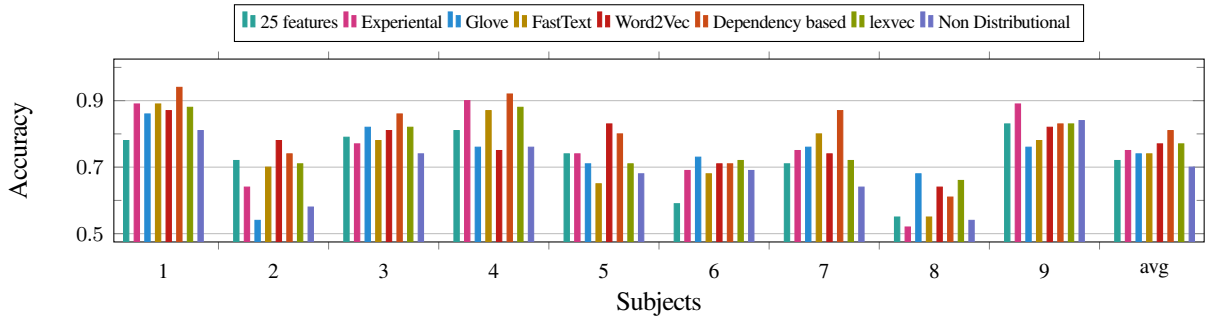


Figure 2: Results of different word representation models for the word to brain activation prediction task for the limited set of word, split per subject.

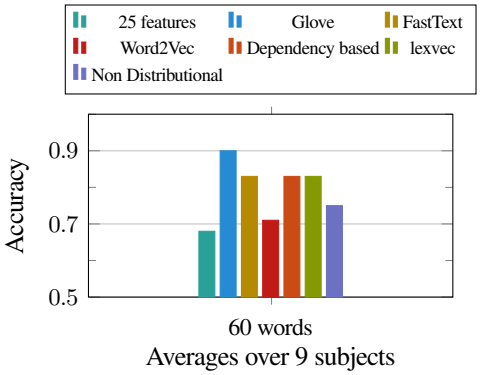


Figure 3: Results of different word representation models for the brain activation to word representation prediction task.

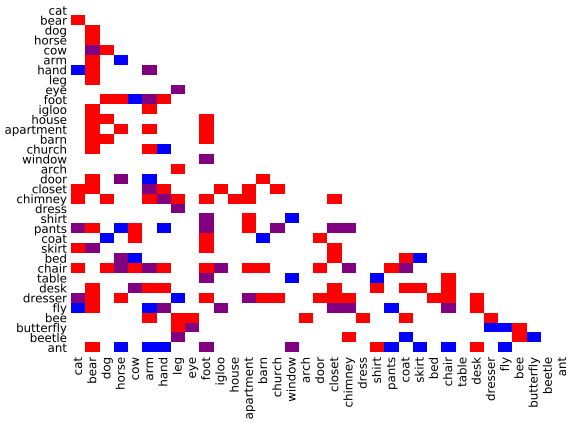


Figure 4: Mismatched word pairs for subject 1: 25 features model (red) vs experiential model (blue). In purple, word pairs confused by both models.

25 features model, the accuracy on the reversed task is much lower. This may be because of the way the feature vector for nouns is distributed in the space in this model. Or it could be that neural activation patterns do not encode all the necessary information to approximate these feature values. This could indicate that while the 25 features model is pretty useful in interpreting brain activation patterns it is not a plausible model to simulate how nouns are represented in the human brain. On the other hand, it seems that it is very easy to construct GloVe word vectors from brain activation patterns; this model achieves an accuracy of 90 percent. In (Sudre et al., 2012) accuracy of 91.19 percent is reported on the similar task on MEG data. GloVe is based on the distributional semantics hypothesis, and it is achieved by learning to predict the global co-occurrence statistics of words in a corpus. Hence, obtaining a high accuracy in the word prediction task using GloVe, supports the fact that the context of the words have a major role in the way we learn the meanings of the words. The important thing to notice is that of course the more information we encode in the word representation the more powerful it becomes in predicting neural activation patterns as far as that information are relevant to some extent. However, this alone doesn't imply that the exact same information is encoded in the neural activation patterns. As we can see in our results, compared to GloVe, it's not that easy to reconstruct the Fasttext and dependency based word vectors from the brain activation patterns. What we can conclude, for now, is that while morphological and dependency information is helpful in learning word representations that are to some extent more similar to the neural representation of nouns in our brain. This information is not explicitly encoded in the brain activation patterns.

In the end, only comparing the accuracy of these models does not reveal much about the differences between them and does not mean that the model with the highest accuracy can replace all the others.

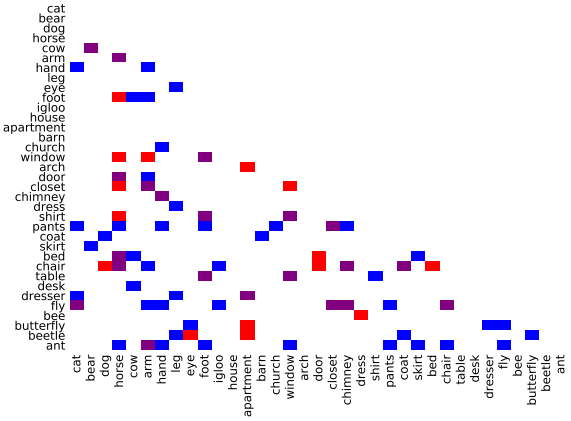


Figure 5: Mismatched pairs for subject 1: dependency based word2vec (red) vs experiential model (blue)

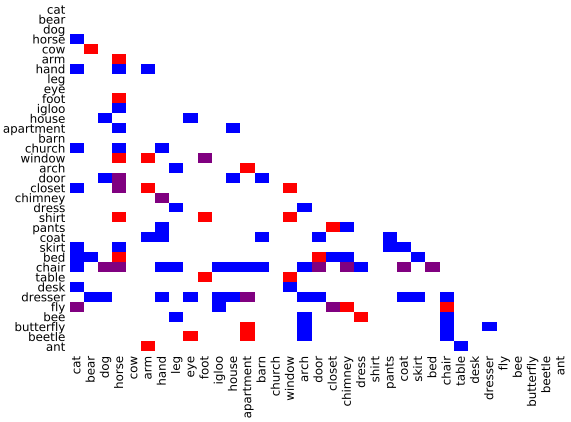


Figure 6: Mismatched pairs for subject 1: dependency based word2vec (red) vs word2vec (blue)

5 *What is the best word embedding model for predicting brain activation for different (classes of) nouns?*

In order to get more insights about the differences between the models, we look into the errors they make. It is informative to see whether each of these models is good at predicting neural activation pattern for a different group of noun pairs. We want to test the hypothesis of whether human brain uses different mechanisms for understanding meanings of different categories of words (Riddoch et al., 1988; Caramazza et al., 1990; Warrington and Shallice, 1984; Caramazza and Shelton, 1998). To investigate this we look into the miss matched noun pairs for each of the word representation models. We want to see which are the most confusing noun pairs for each model and measure the overlap between the errors the models make. This will reveal if these models are actually encoding different kinds of information.

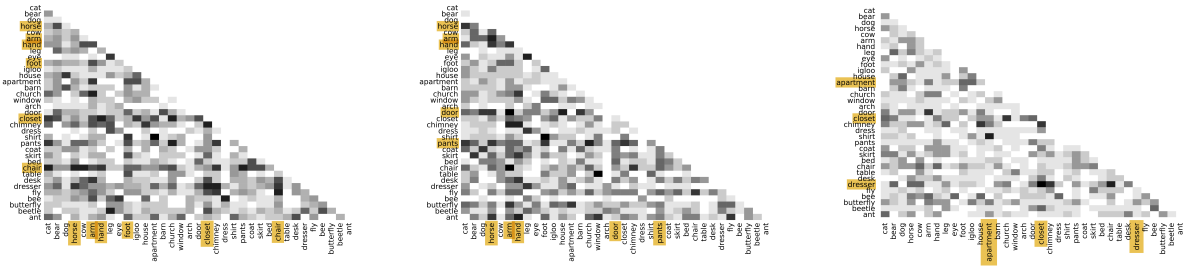
Figures 4, 5 and 6 show the overlap between mismatched pairs for different models for subject 1. In these plots, the red color corresponds to the first model mentioned in the caption, the blue colour corresponds to the second model and the purple colour indicates the overlaps. While there is some overlap between the mistakes of the 25 features model and the experiential model, considerable number of mismatched pairs are not in common between them. One interesting fact about the 25 features model is that for some specific nouns ie. "bear", "foot", "chair", and "dresser", no matter what is its pair, discrimination performance is poor. eg. "bear" is not only confused with other animals, but also with some body parts, places and etc. We do not notice similar phenomena for the experiential model. This could be a side effect of using co-occurrence statistics from corpora to learn word representations and could show that for some reason the representations learned for these nouns are not distinguishable from other nouns. Looking into the noun pair mismatches of the experiential model and the dependency based word2vec in Figure 5, again we see a considerable amount of overlap. They both perform equally for discriminating among animals. But the experiential model makes more mistake about "body parts" and "insects". Comparing the dependency based word2vec with simple word2vec, in Figure 6 we observe similar patterns to Figure 4. As illustrated in the plot, discriminating some words eg. "chair" is difficult for word2vec while it's not the case for dependency based word2vec. It seems like both experiential attributes of nouns and the dependency information is helping in learning more distinguishable representations for nouns.

5.1 25 features vs experiential

As shown in Figure1, the experiential model performs better than the 25 features model in average. Considering the fact that these two models are reflecting the same underlying theory, we might expect that if one of them is more accurate, it can replace the other. However, by looking into the difference between their mismatched pair, Figure 7, we observe that the mistakes these two models make are not completely overlapping: the nouns 'arm' and 'hand' are difficult to discriminate for both models, while 'chair' and 'house' are among the nouns with most mistakes for the 25 features model, and 'horse' and 'door' for the experiential model. For both models, most mismatches are in the category of body parts.

5.2 GloVe vs Dependency-based word2vec

We also compare the mismatch pairs for GloVe and dependency based word2vec as the two neural models

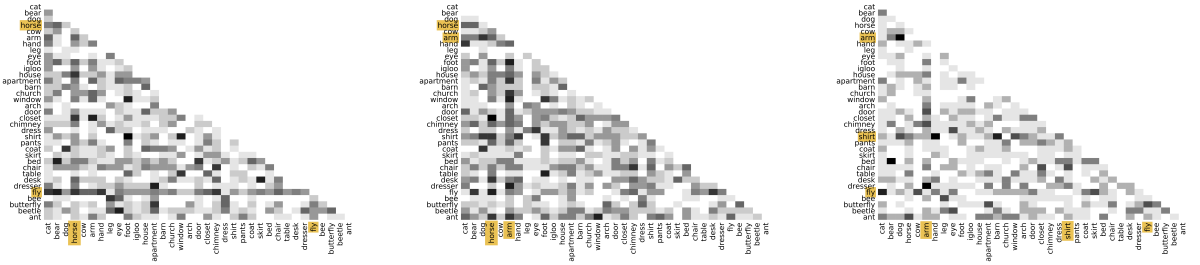


(a) Mismatched pairs for the 25 features model

(b) Mismatched pairs for the experiential model

(c) Difference between mismatched pairs

Figure 7: Comparing mismatched pairs for the 25 features model and the experiential model averaged over all subjects. Axes are the same as in figure 4.



(a) Mismatched pairs for dependency based word2vec

(b) Mismatched pairs for GloVe

(c) Difference between mismatched pairs

Figure 8: Comparing mismatched pairs for dependency based word2vec and GloVe averaged over all subjects

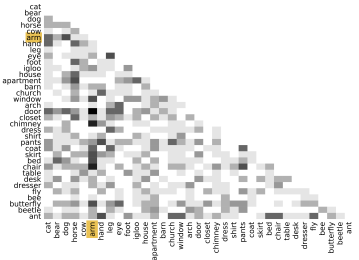


Figure 9: Difference of mismatched pairs for dependency based word2vec and experiential model

that achieve the highest accuracies in Figure 8. These two models are different both in the richness of the information they use to learn word representations, and also the way they use this information. In glove, the model is trained based on the global co-occurance of words whereas in word2vec word representations are learned based on the context of the words for each example locally. For GloVe, similar to the 25 features model and the experiential model, 'arm' is one of the hardest to discriminate nouns. But the 'body parts' category is not as confusing as for the experience based models. For the dependency-based word2vec, the patterns of errors are somehow different and the most difficult word seems to be 'fly'. This is because 'fly' can be either verb and noun, and since it is more frequent as a verb, the dependency-based model is learning the representation of its verb form. For GloVe, this is not very problematic because it is only based on co-occurrence counts, thus an average representation is learned. In general, despite the fact that these two models are based on different assumptions their mismatches have more overlap than for the two experiential models. This may be a side effect of the fact that they both make fewer mistakes.

5.3 Experiential vs Dependency-based word2vec

The mismatched pairs of the experiential model and the dependency based word2vec and their difference is illustrated in Figure 9. The experiential model seems to have less prediction accuracy for noun pairs in the same category.

6 *Which are the most predictable voxels in the brain for each word embedding model?*

Each of the computational models of word representation we have employed to predict brain data is based on modelling different aspects of words meanings. Now we want to investigate if our brain is doing a combination of all these mechanisms and different groups of voxels in the brain are responsible for processing each aspect? One way to test this is to look into the predictability of different voxels with each

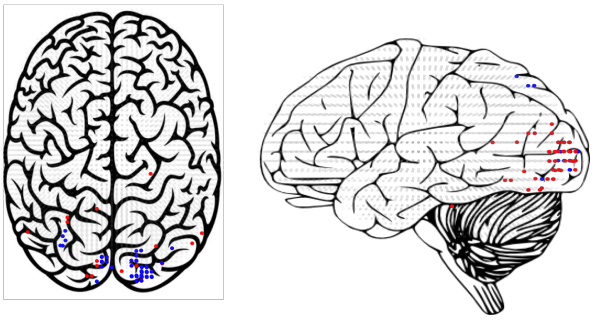


Figure 10: Most predictable voxels for dependecy based word2vec(red) and the experiential model(blue)

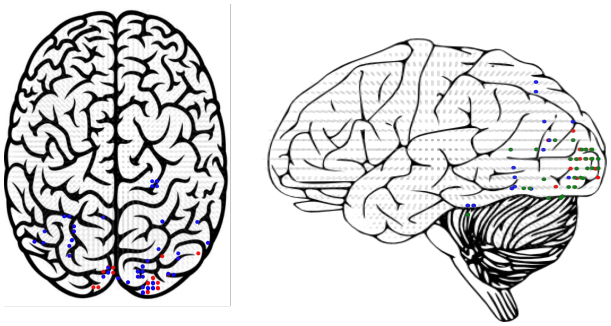


Figure 11: Most predictable voxels for dependecy word2vec(red) and word2vec(blue). Green dots are among the top 50 voxels of both models.

of these models. For this purpose, we have identified the top 50 most predictable voxels for each model. In Figure 10 you can see the 50 most predictable voxels for dependency-based word2vec and the experiential model. In Figure 11 you can see the 50 most predictable voxels for dependency-based word2vec and simple word2vec. The green colour indicates the common top voxels between the two models. From these figures, we can see that there is a lot more overlap between the dependency based word2vec and word2vec, compared to the experiential model.

A Mixed Model If each model is good at predicting the neural activation pattern for a different group of nouns/different groups of voxels, theoretically, it is possible to build a better model using an integrated model. In other words, we should be able to improve the accuracy of predicting neural activation patterns by employing a combined model. We conduct a new experiment by integrating the dependency based word2vec as a neural corpus-based word representation with the experience based models, ie the 25 verbs model and the experiential model. We expect the performance of the model to be a little bit higher than the dependency based word2vec. Our results indicate that combining the dependency based word2vec with the experiential model linearly doesn't lead to an improvement in the accuracy over the limited set of words available in the experiential model. However, linearly combining the 25 feature model with the dependency based word2vec leads to an accuracy of 82 percent over the 60 nouns, which is 2 percent higher than the accuracy of the dependency-based model.

7 Discussion and Conclusion

Based on our systematic comparison, we can conclude that the deep learning models for learning word representations fit very well with brain imaging data. The existing models, like dependency based word2vec, are already beating the experiential word representation models that are particularly designed for the brain activation decoding tasks. Moreover, comparing the results of learning the mappings from words to brain activations and vice versa, convinces us that it is important to study the performance of the models in both directions to really understand what kind of information is encoded in the neural activation patterns for words.

Looking into the details of the performance of these models, it turned out that each of them makes different kinds of mistakes. One of the main problems of the corpus based distributional models that we have applied is that they do not account for different senses of the words. Hence, the representations they learn for words with more than one sense can be noisy and biased toward the most frequent sense. Taking the differences between the models into account, we build a model that combines the experience based word representation model with the dependency based word2vec. By linearly combining the 25 features model with the dependency-based model we are able to achieve a higher accuracy on the brain activation prediction task. We think it is possible to build new models upon the dependency based word2vec which also encode experiential information. One possible approach to achieve this goal is to train word embedding models in a multi-task learning framework with the downstream tasks that reflect different types of real-life experiences in addition to language modelling tasks.

In addition, in order to have a better understanding of the differences between different word representation models, we need to do a further analysis to answer the question *Which are the most predictable voxels in the brain for each word embedding model?*

8 Acknowledgement

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Rasyan Ahmed. 2017. How the brain gives meaning to words. Unpublished Bachelor thesis, Artificial Intelligence, University of Amsterdam.

Andrew J Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics* 5:17–30.

Andrew James Anderson, Jeffrey R Binder, Leonardo Fernandino, Colin J Humphries, Lisa L Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. 2016a. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex* pages 1–17.

Andrew James Anderson, Benjamin D Zinszer, and Rajeev DS Raizada. 2016b. Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities. *NeuroImage* 128:44–53.

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

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 86–90.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*. pages 238–247.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology* 33(3-4):130–174.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* .

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS computational biology* 10(12):e1003963.

Alfonso Caramazza, Argye E Hillis, Brenda C Rapp, and Cristina Romani. 1990. The multiple semantics hypothesis: Multiple confusions? *Cognitive neuropsychology* 7(3):161–189.

Alfonso Caramazza and Jennifer R Shelton. 1998. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of cognitive neuroscience* 10(1):1–34.

Barry Devereux, Colin Kelly, and Anna Korhonen. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*. Association for Computational Linguistics, pages 70–78.

Guy Dove. 2009. Beyond perceptual symbols: A call for representational pluralism. *Cognition* 110(3):412–431.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of ACL*.

Leonardo Fernandino, Colin J Humphries, Mark S Seidenberg, William L Gross, Lisa L Conant, and Jeffrey R Binder. 2015. Predicting brain activation patterns associated with individual lexical concepts based on five sensory-motor attributes. *Neuropsychologia* 76:17–26.

Ahmad Babaeian Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. 2010. WordNet based features for predicting brain activity associated with meanings of nouns. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*. Association for Computational Linguistics, pages 18–26.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. 2014. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology* 10(11):e1003915.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL (2)*. pages 302–308.

Max Mijnheer. 2017. Combining experiential and distributional semantic data to predict neural activity patterns. Unpublished Bachelor thesis, Artificial Intelligence, University of Amsterdam.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

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

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 114–123.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe:global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive neuropsychology* 33(3-4):175–190.

M Jane Riddoch, Glyn W Humphreys, Max Coltheart, and Elaine Funnell. 1988. Semantic systems or system? neuropsychological evidence re-examined. *Cognitive Neuropsychology* 5(1):3–25.

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016a. Enhancing the lexvec distributed word representation model using positional contexts and external memory. *arXiv preprint arXiv:1606.01283* .

Alexandre Salle, Marco Idiart, and Aline Villavicencio. 2016b. Matrix factorization using window sampling and negative sampling for improved word representations. *arXiv preprint arXiv:1606.00819* .

Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage* 62(1):451–463.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann L Cun, and Rob Fergus. 2013. Regularization of neural networks using dropconnect. In *Proceedings of the 30th international conference on machine learning (ICML-13)*. pages 1058–1066.

Elizabeth K Warrington and Tim Shallice. 1984. Category specific semantic impairments. *Brain* 107(3):829–853.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* 9(11):e112575.

Haoyan Xu, Brian Murphy, and Alona Fyshe. 2016. Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2017–2021.

Daniel LK Yamins and James J DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* 19(3):356.