Predicting Brain Activation with WordNet Embeddings

João António Rodrigues, Ruben Branco, João Ricardo Silva, Chakaveh Saedi, António Branco
University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
{joao.rodrigues, ruben.branco, jsilva, chakaveh.saedi, antonio.branco}@di.fc.ul.pt

Abstract

The task of taking a semantic representation of a noun and predicting the brain activity triggered by it in terms of fMRI spatial patterns was pioneered by Mitchell et al. (2008). That seminal work used word co-occurrence features to represent the meaning of the nouns. Even though the task does not impose any specific type of semantic representation, the vast majority of subsequent approaches resort to feature-based models or to semantic spaces (aka word embeddings). We address this task, with competitive results, by using instead a semantic network to encode lexical semantics, thus providing further evidence for the cognitive plausibility of this approach to model lexical meaning.

1 Introduction

Neurosemantics studies the mapping between concepts and the corresponding brain activity, bringing together neuroscientists doing brain imaging research and linguists doing research on the semantics of natural language expressions.

The task introduced by Mitchell et al. (2008) consists of taking a semantic representation of a noun and predicting the functional magnetic resonance imaging (fMRI) spatial activation patterns in the brain triggered by that noun. That is, given a meaning representation of a word, it should be the basis to predict the activation strength at each point (voxel) in the 3D volume of the brain associated to the cognitive handling of that word. This allows to make testable predictions of fMRI activity, even for nouns for which there is no fMRI data available, as long as there is some way to model and represent the semantics of a lexicon.

In lexical semantics, three broad families of approaches have emerged to model meaning, namely (i) semantic networks, (ii) feature-based models, and (iii) semantic spaces. The models of the lexicon produced under these approaches have been embedded in wider models of the whole grammar or in language technology applications and tasks, including synonym identification, analogies detection a.o., were they have been tested on behavioral data sets. The prediction of brain activation considered here is agnostic regarding the approach used to model lexical meaning, thus providing another way of assessing the cognitive plausibility of lexical semantic representations of different sorts.

While most approaches to this task have resorted to feature-based models or to semantic spaces (aka word embeddings), here *we address the task of prediciting the brain activation triggred by nouns rather by using a semantic network, thus providing further evidence for the cognitive plausibility of this approach to model lexical meaning*.

In this paper, we report on the competitive results of resolving the brain activation task by taking a mainstream lexical semantics network, WordNet (Fellbaum, 1998), and resorting to intermediate word embeddings obatined with a novel methodology (Saedi et al., 2018) for generating semantic spaces from semantic networks.

2 The brain activation prediction task

The seminal work of Mitchell et al. (2008) introduced the task consisting of predicting the fMRI activation patterns triggered by a noun-picture pair from a semantic representation of that noun. The language of the data used was English.

Each word w was represented by a set of semantic features given by the normalized co-occurrence counts of w with a set of 25 verbs. These counts were obtained from the Web 1T 5-gram data set

Proceedings of the Eighth Workshop on Cognitive Aspects of Computational Language Learning and Processing, pages 1–5
Melbourne, Australia, July 19, 2018.

(Brants and Franz, 2006), using the n-grams up to length 5 generated from 1 trillion tokens of text.

The 25 verbs were manually selected due to their correspondence to basic sensory and motor activities.[1] Sensory-motor features should be particularly relevant for the representation of objects and, in fact, alternative features based on a random selection of 25 frequent words performed worse.

The fMRI activation pattern at every voxel in the brain is calculated as a weighted sum of each of the 25 semantic features, where the weights are learned by regression to maximum likelihood estimates given observed fMRI data.

To produce the fMRI data, 9 participants were shown 60 different word-picture pairs,[2] the stimuli, each presented 6 times. For each participant, a representative fMRI image for each stimulus was calculated by determining the mean fMRI response from the 6 repetitions and subtracting from each the mean of all 60 stimuli.

Separate models were learned for each of the 9 participants. These models were evaluated using leave-two-out cross-validation, where in each cross-validation iteration the model was asked to predict the fMRI activation for the two held-out words. The two predictions were matched against the two observed activations for those words using cosine similarity over the 500 most stable voxels.

Randomly assigning the two predictions to the two observations would yield a 0.50 accuracy. The models in the seminal paper (Mitchell et al., 2008) achieve a mean accuracy of 0.77, with all individual accuracies significantly above chance.

These results support the plausibility of the two key assumptions underlying the task, namely that (i) brain activation patterns can be predicted from semantic representations of words; and that (ii) lexical semantics can be captured by co-occurrence statistics, the assumption underlying semantics space models of the lexicon.

3 Related work

Several authors have addressed this brain activation prediction task, keeping up with its basic assumptions and resorting to the same data sets for the sake of the comparability of the performance scores obtained.

In an initial period, different authors sought to explore the experimental space of the task by focusing on different ways to set up the features.

Devereux et al. (2010) find that choosing the set of verbs used for the semantic features under an automatic approach can lead to predictions that are equally good as when using the manually selected set of verbs. Jelodar et al. (2010) use the same set of 25 features to represent a word, but instead of basing the features on co-occurrence counts they resort to relatedness measures based on WordNet. Fernandino et al. (2015) use instead a set of features with 5 sensory-motor experience based attributes (sound, color, visual motion, shape, and manipulation). The relatedness scores between the stimulus word and the attributes are based on human ratings instead of corpus data.

Subsequently, as distributional semantics became increasingly popular, authors moved from feature-based representations of the meaning of words to experiment with different vector based representation models (aka word embeddings).

Murphy et al. (2012) compare different corpus-based models to derive word embeddings. They find the best results with dependency-based embeddings, where words inside the context window are extended with grammatical functions. Binder et al. (2016) use word representations based on 65 experiential attributes with relatedness scores crowdsourced from over 1,700 participants. Xu et al. (2016) present BrainBench, a workbench to test embedding models on both behavioral and brain imaging data sets. Anderson et al. (2017) use a linguistic model based on word2vec embeddings and a visual model built with a deep convolutional neural network on the Google Images data set.

Recently, Abnar et al. (2018) evaluated 8 different embeddings regarding their usefulness in predicting neural activation patterns: the co-occurrence embeddings of (Mitchell et al., 2008); the experiential embeddings of (Binder et al., 2016); the non-distributional feature-based embeddings of (Faruqui and Dyer, 2015); and 5 different distributional embeddings, namely word2vec (Mikolov et al., 2013), Fasttext (Bojanowski et al., 2016), dependency-based word2vec (Levy and Goldberg, 2014), GloVe (Pennington et al., 2014) and LexVec (Salle et al., 2016). These authors found that dependency-

[1] The verbs are: *approach*, *break*, *clean*, *drive*, *eat*, *enter*, *fear*, *fill*, *hear*, *lift*, *listen*, *manipulate*, *move*, *near*, *open*, *push*, *ride*, *rub*, *run*, *say*, *see*, *smell*, *taste*, *touch*, and *wear*.

[2] The 60 pairs are composed of 5 items from each of the 12 concrete semantic categories (animals, body parts, buildings, building parts, clothing, furniture, insects, kitchen items, tools, vegetables, vehicles, and other man-made items).

based word2vec achieves the best performance among the approaches resorting to word embeddings, while the seminal approach resorting to 25 features "is doing slightly better on average" with respect to all the approaches experimented with.

The rationale guiding the various works presented in this Section is that the better the performance of the system the higher is the cognitive plausibility of the lexical semantics model resorted to. It is also important to note, however, that there is not always a clearly better method since results show that different methods have different error patterns (Abnar et al., 2018).

4 WordNet embeddings

The previous Sections indicate that approaches to the brain activation task typically resort to feature-based models or to semantic spaces to represent the meaning of words.

In this paper, we address this task by using instead a semantic network as the base repository of lexical semantic knowledge, namely WordNet. We then resort to a novel methodology developed by us (Saedi et al., 2018) for generating semantic space embeddings from semantic networks, and use it to obtain WordNet embeddings. This method is based on the intuition that the larger the number of paths and the shorter the paths connecting any two nodes in a network the stronger is their semantic association.

The conversion method begins by representing the semantic graph as an adjacency matrix M, where element M_{ij} is set to 1 if there is an edge between word w_i and word w_j, and 0 otherwise. Then, this initial relatedness of immediately adjacent words is "propagated" through the matrix by iterating the following cumulative addition

$$M_G^{(n)} = I + \alpha M + \alpha^2 M^2 + \cdots + \alpha^n M^n \quad (1)$$

where I is the identity matrix, the n-th power of the transition matrix, M^n, is the matrix where each M_{ij} counts the number of paths of lenght n between nodes i and j, and α is a decay factor.

The limit of this sum is given by the following closed expression (see Newman, 2010, Eq. 7.63):

$$M_G = \sum_{e=0}^{\infty} (\alpha M)^e = (I - \alpha M)^{-1} \quad (2)$$

Matrix M_G is subsequently submitted to a Positive Point-wise Mutual Information transformation, each line is L2-normalized and, finally, Principal Component Analysis is applied, reducing each line to the size of the desired embedding space. Row i of matrix M_G is then taken as the embedding for word w_i.

Using the methodology outlined above, embeddings with size 850 were extracted from a subset of 60k words in version 3 of English WordNet.[3] When run on the mainstream semantic similarity data set SimLex-999 (Hill et al., 2016), the resulting embeddings showed highly competitive results, outperforming word2vec by some 15% We refer to our embeddings as wnet2vec.[4]

5 Experiment

The good results obtained with wnet2vec in the semantic similarity task lead to experiment with them also in the brain activation prediction task.

5.1 System training

We resorted to the framework implementation[5] by Abnar et al. (2018). Training ran for 1,000 epochs, with a batch size of 29 and a learning rate of 0.001. The loss function is calculated by adding the Huber loss, the mean pairwise squared error and the L2-norm (on weights and bias). Like in previous works, only the 500 most stable voxels are selected. Training was done on a Tesla K40m GPU and took 54 hours (6 hours per subject).

Figure 1 shows an example for Participant 1, with the model prediction and the observed fMRI activation pattern for the word *eye*. The brain activation images were generated with Nibabel (Brett et al., 2017) and Nilearn (Abraham et al., 2014).

5.2 Evaluation and discussion

We followed the usual evaluation procedure for this framework. The cross-validated, leave-two-out mean accuracy was 0.71. The full scores, together with the scores from the original paper, are summarized in Table 1 and shown graphically in Figure 2 (0.50 corresponds to chance).[6]

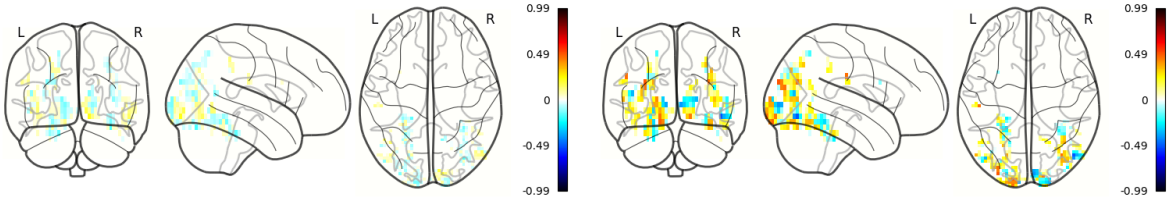
This indicates that wnet2vec has a competitive performance in this task as the mean score obtained is in the range of the scores found for all ap-

[3] We used less than half of the ~150k words in WordNet due to computational limitations as the matrix inverse in (2) faces substantial challenges in terms of the memory footprint.

[4] Available at https://github.com/nlx-group/WordNetEmbeddings

[5] https://github.com/samiraabnar/NeuroSemantics/

[6] Materials for replication available at https://github.com/nlx-group/BrainActivation



(a) predicted (b) observed

Figure 1: fMRI activations for Participant 1, word *eye*

Embeddings	P1	P2	P3	P4	P5	P6	P7	P8	P9	mean
(Mitchell et al., 2008)	0.83	0.76	0.78	0.72	0.78	0.85	0.73	0.68	0.82	0.77
wnet2vec	0.84	0.72	0.86	0.75	0.60	0.67	0.70	0.53	0.74	0.71

Table 1: Accuracy results for the 9 subjects

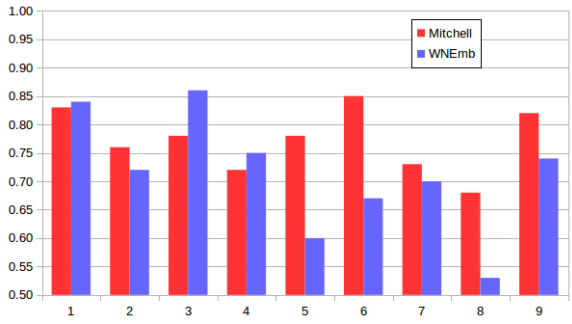


Figure 2: Accuracy results for the 9 subjects

proaches resorting to word embeddings, systematically tested by (Abnar et al., 2018).

In line with all approaches resorting to word embeddings (Abnar et al., 2018), the mean score obtained is also not outperforming the original 25 verb-based co-occurrence features model reported in the seminal paper (Mitchell et al., 2008).

When comparing the scores per participant, the bulk of the wnet2vec losses are due to P5, P6 and P8. For the other subjects, results are close or, in three cases, even better than those from the seminal paper. This highlights the point already made in (Abnar et al., 2018), that different methods have different error patterns, which suggests that an ensemble of classifiers could lead to better overall accuracy. And also, that a dataset with only 9 subjects — the dataset used in the literature on this task since (Mitchell et al., 2008) — may be hindering better empirically grounded conclusions.

Finally, it should be noted that these competitive results were obtained with wnet2vec generated on the basis of 60k words only, thus less than half of WordNet. It will be very interesting to see how the performance of this approach progresses when larger portions of WordNet are taken into account as computational limitations can be overcome.

6 Conclusions

We report on an experiment with the task of predicting the fMRI spatial activation patterns in the brain associated with a given noun.

We resorted to a semantic network of lexical knowledge, viz. WordNet, and thus to a representation of the meaning of the input nouns as elements of concept nodes in a graph of semantically related edges. We also resorted to a derived intermediate vectorial semantic representation (word embeddings) for the input nouns that was obtained by a novel methodology to convert semantic networks into semantic spaces, applied to WordNet.

The results indicate that this model has a competitive performance as its scores are within the range of the results obtained with state of the art models based on corpus-based word embeddings reported in the literature. Though for one third of the 9 subjects this model surpasses Mitchell et al. (2008), on average it did not outperform that seminal model, which used hand-selected features.

The fact that less than half of the words in WordNet were used allows a positive expectation with respect to the strength of the proposed approach, and points towards future work that will seek to use larger portions of WordNet, and further lexical semantics networks and ontologies.

References

Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018) , pages 5766*.

Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and GaelVaroquaux. 2014. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8:14.

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 of Computational Linguistics*, 5(1):17–30.

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*.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1.

Matthew Brett, Michael Hanke, et al. 2017. nipy/nibabel: 2.2.0.

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*, pages 70–78. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2015. Non-distributional word vector representations. *arXiv preprint arXiv:1506.05230*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

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.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41:665–695.

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*, pages 18–26. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, pages 302–308.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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*, pages 114–123. Association for Computational Linguistics.

Mark Newman. 2010. *Networks: An Introduction*. Oxford University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Chakaveh Saedi, Antnio Branco, Joo Antnio Rodrigues, and Joo Ricardo Silva. 2018. Wordnet embeddings. In *Proceedings of the ACL2018 3rd Workshop on Representation Learning for Natural Language Processing (RepL4NLP)*. Association for Computational Linguistics.

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

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 (EMNLP)*, pages 2017–2021.