Understanding the Semantic Space: How Word Meanings Dynamically Adapt in the Context of a Sentence

Nora Aguirre-Celis[1,2] and Risto Miikkulainen[2]

[1] ITESM, E. Garza Sada 2501, Monterrey, NL, 64840, Mexico
[2] The University of Texas in Austin, 2317 Speedway, Austin, TX, 78712 US
{naguirre,risto}@cs.utexas.edu

Abstract

How do people understand the meaning of the word *small* when used to describe a mosquito, a church, or a planet? While humans have a remarkable ability to form meanings by combining existing concepts, modeling this process is challenging. This paper addresses that challenge through CEREBRA (Context-dEpendent meaning REpresentations in the BRAin) neural network model. CEREBRA characterizes how word meanings dynamically adapt in the context of a sentence by decomposing sentence fMRI into words and words into embodied brain-based semantic features. It demonstrates that words in different contexts have different representations and the word meaning changes in a way that is meaningful to human subjects. CEREBRA's context-based representations can potentially be used to make NLP applications more human-like.

1 Introduction

The properties associated with a word such as *small* vary in context-dependent ways: It is necessary to know what the word means, but also the context in which is used, and how the words combine in order to construct the word meaning. Humans have a remarkable ability to form meanings by combining existing concepts. Modeling this process is difficult (Hampton, 1997; Janetzko 2001; Middleton et al, 2011; Murphy, 1988; Pecher et al., 2004; Sag et al., 2001, Wisniewski, 1997, 1998; Yee et al., 2016). How are concepts represented in the brain? How do word meanings change during concept combination or under the context of a sentence? What tools and approaches serve to quantify such changes?

Significant progress has been made in understanding how concepts and word meanings are represented in the brain. In particular, the first two issues are addressed by the Concept Attribute Representation theory (CAR; Binder et al., 2009, 2011, 2016a, 2016b). CAR theory represents concepts as a set of features that constitute the basic components of meaning in terms of known brain systems. It relates semantic content to systematic modulation in neuroimaging activity (fMRI patterns). It suggests that word meanings are instantiated by the weights given to different feature dimensions according to the context. The third issue is addressed by the CEREBRA or Context-dependent mEaning REpresentation in the BRAin neural network model (Aguirre-Celis & Miikkulainen, 2017, 2018, 2019, 2020a, 2020b). It is based on the CAR theory to characterize how the attribute weighting changes across contexts.

In this paper the CAR theory is first reviewed. Then, the CEREBRA model is introduced, followed by the data that provides the basis for the model. Later, experimental results are presented, showing an individual example on the concept combination effect on word meanings, how this effect applies to the entire corpus, and a behavioral analysis to evaluate the neural network model.

2 The CAR Theory

CARs (a.k.a. The Experiential attribute representation model), represent the basic components of meaning defined in terms of neural processes and brain systems. They are composed of a list of well-known modalities that correspond to specialized sensory, motor and affective brain processes, systems processing spatial, temporal, and casual information, and areas involved in social cognition. (Anderson et al., 2016, 2017,

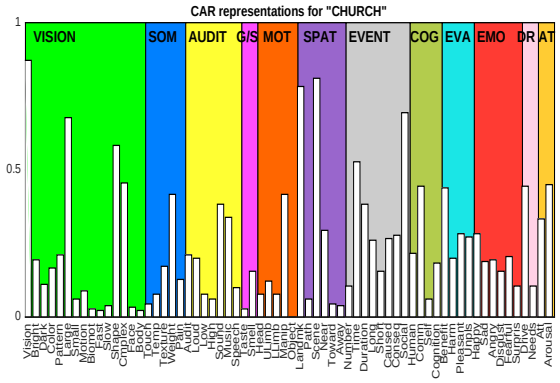


Figure 1: Bar plot of the 66 semantic features for the word *church* (Binder et al., 2009, 2011, 2016a). Given that *church* is an object, it has low weightings on animate attributes such as Face, Body, and Speech, and high weighting on attributes like Vision, Shape, and Weight. However, since it is a building for worship, it does include stronger weightings for spatial attributes such as Landmark and Scene, event attributes like Social, Time and Duration, as well as others such as Communication and Benefit. CAR weighted features for the word *church*.

2018, 2019; Binder et al. 2016a). It is supported by substantial evidence on how humans acquire and learn concepts (Binder et al., 2009, 2011, 2016a, 2016b). The central axiom of this theory is that concept knowledge is built from experience, as a result, knowledge representation in the brain is dynamic.

The features are weighted according to statistical regularities. The semantic content of a given concept is estimated from ratings provided by human participants. For example, concepts referring to things that make sounds (e.g., *explosion*, *thunder*) receive high ratings on a feature representing auditory experience, relative to things that do not make a sound (e.g., *milk*, *flower*).

Each word is modeled as a collection of 66 features that captures the strength of association between each neural attribute and word meaning. Specifically, the degree of activation of each attribute associated with the concept can be modified depending on the linguistic context, or combination of words in which the concept occurs. More detailed account of the attribute selection and definition is given by Binder, et al. (2009, 2011, 2016a, and 2016b).

Figure 1, shows an example of the weighted CARs for the concept *church*. The weight values represent average human ratings for each feature. Given that *church* is an object, it has low

Terminology

CARWord: The neural network input. CARWords are formed based on ratings by human subjects (Section 3.3). They are the original brain-based semantic representations of words, i.e., word without context. Each CARWord is a vector of 66 attributes.

CARWordRevised: The input of the neural network after FGREP. CARWordsRevised are formed by FGREP modifying the original CARWords. They are the context-dependent meaning representations of words for each sentence where they occurred. Each CARWordRevised is a vector of 66 attributes.

ε : The error signal. The SynthSent is subtracted voxelwise from the fMRISent to produce an error signal. Each error is a vector of 396 changes.

fMRISent: The neural network target. They are the original brain data collected from human subjects using neuroimaging. Each fMRISent is a vector of 396 voxels.

SyntSent: The predicted fMRI sentence after training. The SynthWords in the sentence are averaged to form this prediction. Each SynthSent is a vector of 396 values.

SyntSentRevised: The modified SyntSent after applying the error signal changes. Each of these SynthSentRevised is a vector of 396 values.

SyntWord: The neural network target. They are derived by averaging the fMRISent. They are synthetic because individual fMRI data for words do not exist, thus they are obtained by averaging each fMRISent where the word occurred. Each SynthWord is a vector of 396 voxels.

SyntWordRevised: The target for the neural network after FGREP. They are derived from the SynthSentRevised using the error signal changes.

W1..W3: labels for each CARWord in a sentence.

W'1..W'3: labels for each SynthWord in a sentence.

Figure 2: Terminology for the abbreviated terms used in the CEREBRA model.

weightings on animate attributes such as Face, Body, and Speech, and high weighting on attributes like Vision, Size, Shape, and Weight. However, since it is a building and a place for worship, it does include strong weightings for Sound and Music, spatial attributes such as Landmark and Scene, event attributes like Social, Time and Duration, as well as others such as Communication and Benefit.

3 The CEREBRA Model

Building on the idea of grounded word representation in CAR theory, this work aims to understand how word meanings change depending on context. The following sections describe the computational model that characterizes such representations. The specific terms to the CEREBRA model are denoted by abbreviations throughout the paper (e.g., CARWord, fMRISent, SynthWord). For reference, they are described in Figure 2.

3.1 System Design

The overall design of CEREBRA is shown in Figure 3. It is a neural network model that performs two main tasks: Prediction and Interpretation. During the Prediction task, the model form a predicted fMRI for each sentence without the context effects. Each sentence is thus compared against the observed fMRI sentence to calculate an

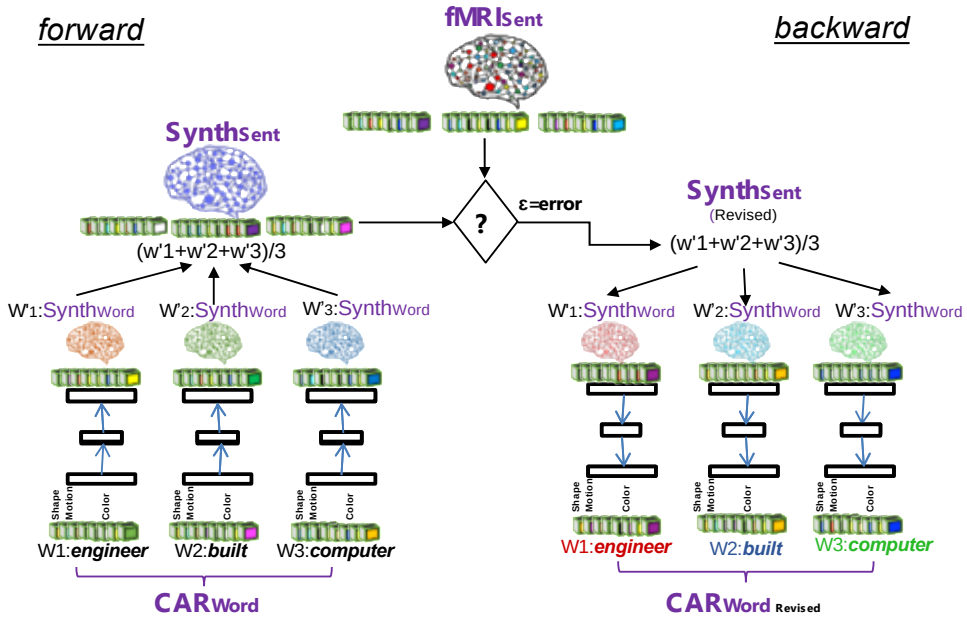


Figure 3: The CEREBRA model to account for context effects. (1) Propagate CARWords to SynthWords. (2) Construct SynthSent by averaging the SynthWords into a prediction of the sentence. (3) Compare SynthSent with the observed fMRI. (4) Backpropagate the error with FGREP for each sentence, freezing network weights and changing only CARWords. (5) Repeat until error reaches zero or CAR components reach their upper or lower limits. The modified CARs represent the word meanings in context. Thus, CEREBRA captures context effects by mapping brain-based semantic representations to fMRI sentence images.

error signal. This error signal is used repeatedly by the Interpretation task. During the Interpretation task, the model is used to determine how the CARs should adjust to eliminate the remaining error. The error is used to change the CARs themselves using the FGREP mechanism (Forming Global Representations with Extended BP, Miikkulainen & Dyer, 1991). The process iterates until the error goes to zero.

3.2 Mapping CARs to Synthetic Words

The CEREBRA model is first trained to map the CARWord representations in each sentence to SynthWords (The "forward" side of Figure 3). It uses a standard three-layer backpropagation neural network (BPNN). Gradient descent is performed for each word, changing the connection weights of the network to learn this task (Rumelharth, et al., 1986).

The BPNN was trained for each of the eleven fMRI subjects for a total of 20 repetitions each, using different random seeds. Complete training thus yields 20 different networks for each subject, resulting in 20 sets of 786 predicted SynthWord representations, that is, one word representation for each sentence where the word appears.

3.3 Sentence Prediction to Change CARs

For the Prediction task, the sentences are assembled using the predicted SynthWords by averaging all the words that occur in the sentence, yielding the prediction sentence called SynthSent. For the Interpretation task, in addition to the construction of the predicted sentence, further steps are required. First, the prediction error is calculated by subtracting the newly constructed predicted SynthSent from the original fMRISent. Then, the error is backpropagated to the inputs CARWords for each sentence (The "backward" side of Figure 3). However, following the FGREP method the weights of the network no longer change. Instead, the error is used to adjust the CARWords in order for the prediction to become accurate.

This process is performed until the prediction error is very small (near zero) or cannot be modified (CARWords already met their limits, i.e., 0 or 1), which is possible since FGREP is run separately for each sentence. These steps are repeated 20 times for each subject. At the end, the average of the 20 representations is used to represent each of the 786 context-based words (CARWord Revised), for each subject.

Eventually, the Revised CARWord represents the word meaning for the current sentence such that, when combined with other Revised CARWords in the sentence, the estimate of sentence fMRI becomes correct.

4 Data Collection and Processing

The CEREBRA model is based on the following sets of data: A sentence collection prepared by Glasgow et al. (2016), the semantic vectors (CAR ratings) for the words obtained via Mechanical Turk, and the fMRI images for the sentences, collected by the Medical College of Wisconsin (Anderson et al., 2016, 2017, 2018, 2019; Binder et al., 2016a, 2016b). Additionally, fMRI representations for individual words (called SynthWord) were synthesized by averaging the sentence fMRI.

4.1 Sentence Collection

A total of 240 sentences were composed of two to five content words from a set of 242 words (141 nouns, 39 adjectives and 62 verbs). The words were selected toward imaginable and concrete objects, actions, settings, roles, state and emotions, and events. Examples of words include *doctor*, *car*, *hospital*, *yellow*, *flood*, *damaged*, *drank*, *accident*, *summer*, *chicken*, and *family*. An example of a sentence containing some of these words is *The accident damaged the yellow car*.

4.2 Semantic Word Vectors

In a separate study Binder et al. (2016a, 2016b) collected CAR ratings for the original 242 words through Amazon Mechanical Turk. In a scale of 0-6, the participants were asked to assign the degree to which a given concept is associated with a specific type of neural component of experience (e.g. "To what degree do you think of a *church* as having a fixed location, as on a map?").

Approximately 30 ratings were collected for each word. After averaging all ratings and removing outliers, the final attributes were transformed to unit length yielding a 66-dimensional feature vector such as the one shown in Figure 1 for the word *church*. Note that this semantic feature approach builds its vector representations by mapping the conceptual content of a word (expressed in the questions) to the corresponding neural systems for which the CAR dimensions stand. This approach thus contrasts with systems where the features are extracted from text corpora and word co-occurrence with no direct association to perceptual grounding (Baroni et. al., 2010; Burgess, 1998; Harris, 1970; Landauer & Dumais, 1997; Mikolov et al., 2013).

4.3 Neural fMRI Sentence Representations

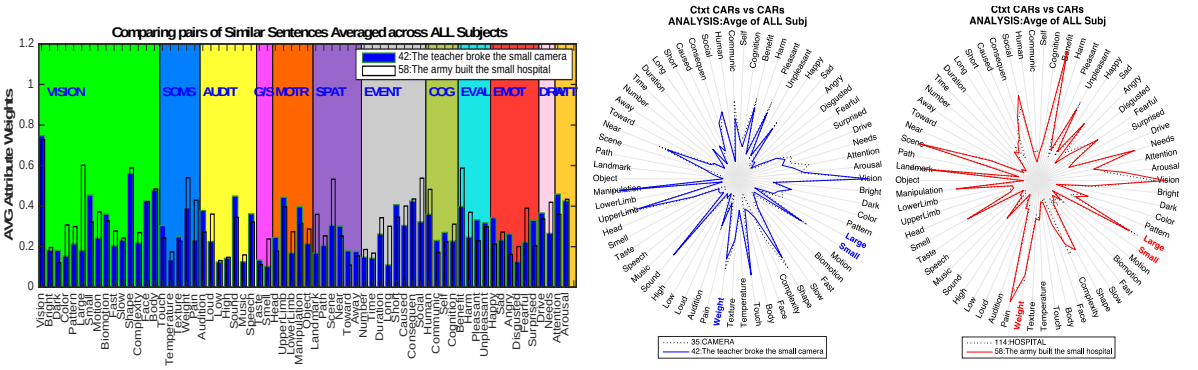
If indeed word meaning changes depending on context, it should be possible to see such changes by directly observing brain activity during word and sentence comprehension. Binder and his team collected twelve repetitions of brain imaging data from eleven subjects by recording visual, sensory, motor, affective, and other brain systems.

To obtain the neural correlates of the 240 sentences, subjects viewed each sentence on a computer screen while in the fMRI scanner. The fMRI patterns were acquired with a whole-body Three-Tesla GE 750 scanner at the Center for Imaging Research of the Medical College of Wisconsin (Anderson, et al., 2016). The sentences were presented word-by-word using a rapid serial visual presentation paradigm, with each content word exposed for 400ms followed by a 200ms inter-stimulus interval. Participants were instructed to read the sentences and think about their overall meaning.

The fMRI data were pre-processed using standard methods. The transformed brain activation patterns were converted into a single-sentence fMRI representation per participant by taking the voxel-wise mean of all repetitions (Anderson et al., 2016; Binder et al., 2016a, 2016b). To form the target for the neural network, the most significant 396 voxels per sentence were then chosen. The size selection mimics six case-role slots of content words consisting of 66 attributes each. The voxels were further scaled to [0.2..0.8].

4.4 Synthetic fMRI Word Representations

The Mapping CARs task in CEREBRA (described in Section 3.2) requires fMRI images for words in isolation. Unfortunately, the collected neural data set does not include such images. Therefore, a technique developed by Anderson et al. (2016) was adopted to approximate them. The voxel values for a word were obtained by averaging all fMRI images for the sentences where the word occurs. These vectors, called SynthWords, encode a combination of examples of that word along with other words that appear in the same sentence. Thus,



(a) Averaged sentences across subjects

(b) Averaged concepts across subjects

Figure 4: The effect of centrality on two contexts for the word *small*. (a) The average for all subjects for the two sentences. (b) The new *camera* and *hospital* representations averaged for all subjects. In the left side of the figure, the new CARs for Sentence 42 have salient activations for an object, denoting the *camera* properties like Dark, Small, Manipulation, Head, Upper Limb, Communication, and emotions such as Sad (e.g., broke the camera). The new CARs for Sentence 58, has high feature activations for large buildings describing a Large, and Heavy structure such as a *hospital*. In the right side of the figure, for each word the central attributes are highlighted to emphasize how same dimensions are more important to some concepts than others. The centrality effect correlation analysis (Medin & Shoben, 1988).

the SynthWord representation for *mouse* obtained from Sentence 56:*The mouse ran into the forest* and Sentence 60:*The man saw the dead mouse* includes aspects of running, forest, man, seeing, and dead, altogether. This process of combining contextual information is similar to several semantic models in computational linguistics (Baroni et al., 2010; Burgess, 1998; Landauer et al., 1997; Mitchell & Lapata, 2010). Additionally, in other studies, this approach has been used successfully to predict brain activation (Anderson et al., 2016, 2017, 2018, 2019; Binder, et al., 2016a, 2016b; Just, et al., 2017).

Due to the limited number of sentences, some of SynthWords became identical and were excluded from the dataset. The final collection includes 237 sentences and 236 words (138 nouns, 38 adjectives and 60 verbs). Similarly, due to noise inherent in the neural data, only eight subject fMRI patterns were used for this study.

5 Experiments

CEREBRA's context-based representations were evaluated through several computational experiments as well as through a behavioral analysis. The computational experiments quantify how the CAR representation of a word changes in different sentences for individual cases by correlating these changes to the CAR representations of the other words in the sentence (OWS). The behavioral study evaluates the CEREBRA context-based representations against human judgements.

5.1 Analysis of an Individual Example

Earlier work showed that (1) words in different contexts have different representations, and (2) these differences are determined by context. These effects were demonstrated by analyzing individual sentence cases across multiple fMRI subjects (Aguirre-Celis & Miikkulainen, 2017, 2018).

Particularly, in this example the attributes of the adjective-noun combinations are analyzed on the centrality effect for the word *small*, as expressed in Sentence 42: *The teacher broke the small camera*, and Sentence 58: *The army built the small hospital*. Centrality expresses the idea that some attributes are true to many different concepts but they are more important to some concepts than others (Medin & Shoben, 1988). For example, it is more important for boomerangs to be curved than for bananas.

Figure 4 shows the differences for *small* in these two contexts. The left side displays all 66 attributes for the two sentence representations averaged across subjects, and the right side displays the context-based representations averaged across all subjects for *camera* and *hospital*.

The size dimensions (e.g., Small and Large), demonstrated the centrality principle for these specific contexts. The left side of Figure 4 shows Sentence 42 (e.g., *small camera*) with salient activation for the central attribute Small and low

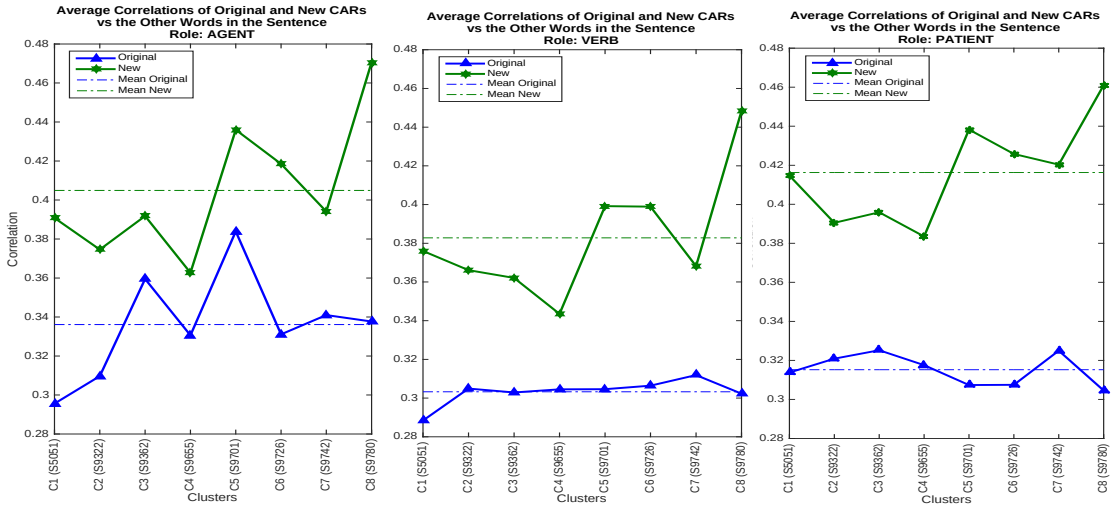


Figure 5: Correlation results per subject cluster and word roles. Average correlations analyzed by word class for eight subjects comparing original and new CARs vs. the average of the OWS respectively. A moderate to strong positive correlation was found between new CARs and the OWS, suggesting that features of one word are transferred to OWS during conceptual combination. Interestingly, the original and new patterns are most similar in the AGENT panel, suggesting that this role encodes much of the context. The results show that the effect occurs consistently across subjects and sentences.

activation for the non-central attribute Large. In contrast, Sentence 57 (e.g., *small hospital*) presents low activation on the non-central attribute Small but high activation on the central attribute Large.

These findings suggest that these attributes are essential to small objects and big structures, respectively. However, the size dimension alone cannot represent the centrality effect completely.

Additionally, given that both *camera* and *hospital* are inanimate objects, the right side of Figure 4 shows that they share low weightings on human-related attributes like Biomotion, Face, Body, and Speech. However, they also differ in expected ways, including salient activations on Darkness, Color, Small and Large size, and Weight. As part of the sentence context, the activations include human-like attributes such as Social, Human, Communication, Pleasant, Happy, Sad and Fearful. Overall, each sentence representation moves towards their respective sentence context (e.g., *camera* or *hospital*).

5.2 Aggregation Analysis

Further work verified the above conclusions in the aggregate through a statistical analysis across an entire corpus of sentences. The goal was to measure how the CARs of a word changes in different sentences, and to correlate these changes to the CARs of the other words in the sentence (OWS). In other words, the conceptual combination effect was quantified statistically across sentences and subjects (Aguirre-Celis & Miikkulainen, 2019, 2020b).

The hypothesis is based on the idea that similar sentences have a similar effect, and this effect is consistent across all words in the sentence. In order to test this hypothesis it is necessary to (1) form clusters of similar sentences for the entire collection, and (2) calculate the average changes on the words identified by the role they play for the same cluster of sentences. Through correlations, it is possible to demonstrate how similar sentences cause analogous changes in words that play identical roles in those sentences.

The results are shown in Figure 5. The correlations are significantly higher for new CARs than for the original CARs across all subjects and all roles. Furthermore, the AGENT role represents a large part of the context in both analyses (i.e., modified and original CARs). Thus, the results confirm that the conceptual combination effect occurs reliably across subjects and sentences, and it is possible to quantify it by analyzing the fMRI images using the CEREBRA model on CARs. As a summary, the average correlation was 0.3201 (stdev 0.020) for original CAR representations and 0.3918 (stdev 0.034) for new CAR representations.

Thus, this process demonstrated that changes in a target word CAR originate from the OWS. For instance, if the OWS have high values in the CAR

HUMAN RESPONSES DISTRIBUTION						
Resp/Part	P1	P2	P3	P4	AVG	%
-1	2065	995	645	1185	1223	34.0%
0	149	1120	1895	1270	1109	30.8%
1	1386	1485	1060	1145	1269	35.3%
TOT	3600	3600	3600	3600	3600	100%
PARTICIPANT AGREEMENT ANALYSIS						
	P1	P2	P3	P4	AVERAGE	%
P1	0	1726	1308	1650	1561	43%
P2	1726	0	1944	1758	1809	50%
P3	1308	1944	0	1741	1664	46%
P4	1650	1758	1741	0	1716	48%
				TOTAL	6751	
				AVG xPAR	1688	
		AVERAGE	Particip match each other			47%

(a) Human Responses

PARTICIPANTS AVERAGE AGREEMENT			
RATINGS	HUMAN	CEREBRA	CHANCE
-1	618	463	8
0	456	3	0
1	892	587	886
TOTAL	1966	1053	894
AVERAGE		54%	45%

(b) Matching Predictions

SUBJECTS	CEREBRA		CHANCE		p-value
	MEAN	VAR	MEAN	VAR	
S5051	1033	707.25	894	6.01	3.92E-24
S9322	1035	233.91	894	7.21	6.10E-33
S9362	1063	224.41	894	11.52	5.22E-36
S9655	1077	94.79	894	7.21	3.89E-44
S9701	1048	252.79	895	12.03	1.83E-33
S9726	1048	205.82	894	4.62	1.73E-35
S9742	1075	216.77	895	7.21	1.65E-37
S9780	1039	366.06	894	2.52	6.10E-30

(c) Statistical Significance

Table 1: Comparing CEREBRA predictions with human judgements. (a) Distribution analysis and inter-rater agreement. The top table shows human judgement distribution for the three responses "less" (-1), "neutral" (0), and "more" (1). The bottom table shows percentage agreement for the four participants. Humans agree 47% of the time. (b) Matching CEREBRA predictions with human data, compared to chance baseline. The table shows the average agreement of the 20 repetitions across all subjects. CEREBRA agrees with human responses 54% while baseline is 45% - which is equivalent to always guessing "more", i.e., the largest category of human responses. (c) Statistical analysis for CEREBRA and baseline. The table shows the means and variances of CEREBRA and chance models for each subject and the p-values of the t-test, showing that the differences are highly significant. Thus, the context-dependent changes are actionable knowledge that can be used to predict human judgements.

spatial dimension for Path, then that dimension in the modified CAR should be higher than in the original CAR, for such target word. The CEREBRA model encodes this effect into the CARs where it can be measured.

5.3 Behavioral Study

While Sections 5.1 and 5.2 showed that differences in the fMRI patterns in sentence reading can be explained by context-dependent changes in the semantic feature representations of the words. The goal of this section is to show that these changes are meaningful to humans. Therefore, human judgements were compared against CEREBRA predictions (Aguirre-Celis & Miikkulainen, 2020a, 2020b).

Measuring Human Judgements: A survey was designed to characterize context-dependent changes by asking the subject directly: In this context, how does this attribute change? Human judgements were crowdsourced using Google Forms. The complete survey was an array of 24 questionnaires that included 15 sentences each. For each sentence, the survey measured 10 attribute changes for each target word. Only the top 10 statistically most significant attribute changes for each target words (roles) were used. Overall, each questionnaire thus contained 150 evaluations. The 24 questionnaires can be found at: https://drive.google.com/drive/folders/1jDCqKMuH-SyTxcJ7oJRbr7mYV6WNNEWH?usp=sharing

Human responses were first characterized through data distribution analysis. Table 1 (a) shows the number of answers "less" (-1), "neutral" (0), and "more" (1) for each participant. Columns labeled P1, P2, P3, and P4 show the answers of the participants. The top part of Table 1 (a) shows the distribution of the raters' responses and the bottom part shows the level of agreement among them. As can be seen from the table, the participants agreed only 47% of the time. Since the inter-rater reliability is too low, only questions that were the most reliable were included, i.e., where three out of four participants agreed. There were 1966 such questions, or 55% of the total set of questions.

Measuring CEREBRA's Predictions: The survey directly asks for the direction of change of a specific word attribute in a particular sentence, compared to the word's generic meaning. Since the changes in the CEREBRA model range within (-1,1), in principle that is exactly what the model produces. However, during the experiments it was found that some word attributes always increase, and do so more in some contexts than others. This

effect is well known in conceptual combination (Hampton, 1997; Wisniewsky, 1998), contextual modulation (Barclay, 1974, Barsalou et al., 1987, 1993), and attribute centrality (Medin & Shoben, 1988). The direction of change is therefore not a good predictor of human responses.

These changes need to be measured relative to changes in the OWS. Thus, the approach was based on asking: What is the effect of CARs used in context as opposed to CARs used in isolation? This effect was measured by computing the average of the CEREBRA changes (i.e., new minus original) of the different representations of the same word in several contexts, and subtracting that average change from the change of the target word.

Matching CEREBRA's Predictions with Human Judgements: In order to demonstrate that the CEREBRA model has captured human performance, the agreements of the CEREBRA changes and human surveys need to be at least above chance. Therefore a baseline model that generated random responses from the distribution of human responses was created. The results are reported in Table 1 (b), and the statistical significance of the comparisons in Table 1 (c).

The CEREBRA model matches human responses in 54% of the questions when the baseline is 45% - which is equivalent to always guessing "more", i.e., the largest category of human responses. The differences shown in Table 1 (c) are highly statistically significant for the eight subjects. These results show that the changes in word meanings (i.e., due to sentence context observed in the fMRI and interpreted by CEREBRA) are real and meaningful to humans (Aguirre-Celis & Miikkulainen, 2020a, 2020b).

6 Discussion and Future Work

This paper described how the CAR theory, the fMRI images, and the CEREBRA model form the groundwork to characterize dynamic word meanings. The CEREBRA model generates good interpretations of word meanings considering that the dataset was limited and was not originally designed to address the dynamic effects in meaning. In future work, it would be interesting to replicate the studies on a more extensive data set. A fully balanced stimuli including sentences with identical contexts (e.g., *The yellow bird flew over the field* vs. *The yellow plane flew over the field*) and contrasting contexts (e.g., *The aggressive dog chased the boy* vs. *The friendly dog chased the boy*), could help characterize the effects in more detail. The context-based changes should be even stronger, and it should be possible to uncover more refined effects. Such data should also improve the survey design, since it would be possible to identify questions where the effects can be expected to be more reliable.

Similarly, it would be desirable to extend the fMRI data with images for individual words. The CEREBRA process of mapping semantic CARs to SynthWords and further to sentence fMRI refines the synthetic representations by removing noise. However, such representations blend together the meanings of many words in many sentences. Thus, by acquiring actual word fMRI, the observed effects should become even more clear.

One disadvantage on CEREBRA is that it is expensive to collect fMRI patterns and human ratings at a massive scale compared to running a statistical algorithm on a data repository. Furthermore, any changes to the model (e.g., adding features) would require new data to be collected. On the other hand, such data provides a grounding to neural processes and behavior that does not exist with statistical approaches.

Concept representation in the CAR approach can be compared to other methods such as Conceptual Spaces (CS; Gardenfors, 2004; Bechberger & Kuhnberger, 2019), and distributional semantic models (DSMs; Anderson et. al., 2013; Bruni et al., 2014; Burgess, 1998; Landauer & Dumais, 1997; Mikolov et al., 2013; Mitchell & Lapata, 2010; Silberer & Lapata, 2014). The CAR theory and CS characterize concepts with a list of features or dimensions as the building blocks. The CAR theory provides a set of primitive features for the analysis of conceptual content in terms of neural processes (grounded in perception and action). The CS framework suggests a set of "quality" dimensions as relations that represent cognitive similarities between stimuli (observations or instances of concepts). CS is also considered a grounding mechanism that connects abstract symbols to the real world. The CAR and CS approaches include similar dimensions (i.e., weight, temperature, brightness) and some of those dimensions are part of a larger domain (e.g., color) or a process (e.g., visual system). Whereas CAR theory is a brain-based semantic representation where people weigh concept dimensions differently based in context,

DSMs are not grounded on perception and motor mechanisms. Instead, DSM representations reflect semantic knowledge acquired through a lifetime of linguistic experience based on statistical co-occurrence. DSMs do not provide precise information about the experienced features of the concept itself (Anderson et al., 2016). In CEREBRA, this grounding provides a multimodal approach where features directly relate semantic content to neural activity.

7 Conclusions

The CEREBRA model was constructed to test the hypothesis that word meanings change dynamically based on context. The results suggest three significant conclusions: (1) context-dependent meaning representations are embedded in the fMRI sentences, (2) they can be characterized using CARs together with the CEREBRA model, and (3) the attribute weighting changes are real and meaningful to human subjects. Thus, CEREBRA opens the door for cognitive scientists to achieve better understanding and form new hypotheses about how semantic knowledge is represented in the brain. Additionally, the context-based representations produced by the model could be used for a broad range of artificial natural language processing systems, where grounding concepts as well as understanding novel combinations of concepts is critical.

Acknowledgments

We would like to thank J. Binder (Wisconsin), R. Raizada and A. Anderson (Rochester), M. Aguilar and P. Connolly (Teledyne) for providing this data and for their valuable help regarding this research. This work was supported in part by IARPA-FA8650-14-C-7357 and by NIH 1U01DC014922 grants.

References

Nora Aguirre-Celis & Risto Miikkulainen. (2017). From Words to Sentences & Back: Characterizing Context-dependent Meaning Representations in the Brain. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK, pp. 1513-1518.

Nora Aguirre-Celis & Risto Miikkulainen. (2018) Combining fMRI Data and Neural Networks to Quantify Contextual Effects in the Brain. In: Wang S. et al. (Eds.). *Brain Informatics*. BI 2018. Lecture Notes in Computer Science. 11309, pp. 129-140. Springer, Cham.

Nora Aguirre-Celis & Risto Miikkulainen. (2019). Quantifying the Conceptual Combination Effect on Words Meanings. *Proceedings of the 41th Annual Conference of the Cognitive Science Society*, Montreal, CA. 1324-1331.

Nora Aguirre-Celis & Risto Miikkulainen. (2020a). Characterizing the Effect of Sentence Context on Word Meanings: Mapping Brain to Behavior. *Computation and Language*. arXiv:2007.13840.

Nora Aguirre-Celis & Risto Miikkulainen. (2020b). Characterizing Dynamic Word Meaning Representations in the Brain. *In Proceedings of the 6th Workshop on Cognitive Aspects of the Lexicon (CogALex-VI)*, Barcelona, ES, December 2020.

Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013); Seattle, WA: Association for Computational Linguistics. pp. 1960–1970.

Andrew J. Anderson, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, Rajeev D S Raizada. 2016. Perdicting Neural activity patterns associated with sentences using neurobiologically motivated model of semantic representation. *Cerebral Cortex*, pp. 1-17. DOI:10.1093/cercor/bhw240

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. *Transaction of the Association for Computational Linguistics* 5: 17-30.

Andrew J. Anderson, Edmund C. Lalor, Feng Lin, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D.S. Raizada, Scott Grimm, and Xixi Wang. 2018. Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. *Cerebral Cortex*, pp. 1-16. DOI:10.1093/cercor/bhy110.

Andrew J. Anderson, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D.S. Raizada, Feng Lin, and Edmund C. Lalor. 2019. An integrated neural decoder of linguistic and experiential meaning. *The Journal of neuroscience: the official journal of the Society for Neuroscience*.

Richard Barclay, John D. Bransford, Jeffery J. Franks, Nancy S. McCarrell, & Kathy Nitsch. 1974.

Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior, 13*:471–481.

Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, 34(2):222-254.

Lawrence W. Barsalou. 1987. The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization.* Cambridge, England: Cambridge University Press.

Lawrence W. Barsalou, Wenchi Yeh, Barbara J. Luka, Karen L. Olseth, Kelly S. Mix, Ling-Ling Wu. 1993. Concepts and Meaning. *Chicago Linguistic Society 29: Papers From the Parasession on Conceptual Representations*, 23-61. University of Chicago.

Lucas Bechberger, Kai-Uwe Kuhnberger. 2019. A Thorough Formalization of Conceptual Spaces. In: Kern-Isberner, G., Furnkranz, J., Thimm, M. (eds.) KI 2017: *Advances in Artificial Intelligence: 40th Annual German Conference on AI*, Dortmund, Germany.

Jeffrey R. Binder and Rutvik H. Desai, William W. Graves, Lisa L. Conant. 2009. Where is the semantic system? A critical review of 120 neuroimaging studies. *Cerebral Cortex*, 19:2767-2769.

Jeffrey R. Binder and Rutvik H. Desai. 2011. The neurobiology of semantic memory. *Trends Cognitive Sciences*, 15(11):527-536.

Jeffrey R. Binder. 2016a. In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, 23. doi:10.3758/s13423-015-0909-1

Jeffrey R. Binder, Lisa L. Conant, Colin J. Humpries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, Rutvik H. Desai. 2016b. Toward a brain-based Componential Semantic Representation. *Cognitive Neuropsychology*, 33(3-4):130-174.

Elia Bruni, Nam Khanh Tran, Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49:1-47.

Curt Burgess. 1998. From simple associations to the building blocks of language: Modeling meaning with HAL. *Behavior Research Methods, Instruments, & Computers*, 30:188–198.

Peter Gardenfors. 2004. Conceptual spaces: The geometry of thought, The MIT Press.

Kimberly Glasgow, Matthew Roos, Amy J. Haufler, Mark Chevillet, Michael Wolmetz. 2016. Evaluating semantic models with word-sentence relatedness. *Computing Research Repository*, arXiv:1603.07253.

James Hampton. 1997. Conceptual combination. In K. Lamberts & D. R. Shanks (Eds.), *Studies in cognition. Knowledge, concepts and categories*, 133–159. MIT Press.

Zellig Harris. 1970. Distributional Structure. *In Papers in Structure and Transformational Linguistics*, 775-794.

Dietmar Janetzko. 2001. Conceptual Combination as Theory Formation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 23.

Marcel A. Just, Jing Wang, Vladimir L. Cherkassky. 2017. Neural representations of the concepts in simple sentences: concept activation prediction and context effect. Neuroimage, 157:511–520.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory. *Psychological Review*, 104:211-240.

Douglas L. Medin and Edward J. Shoben. 1988. Context and structure in conceptual combination. *Cognitive Psychology*, 20:158-190.

Erica L. Middleton, Katherine A. Rawson, and Edward J. Wisniewski. 2011. "How do we process novel conceptual combinations in context?". *Quarterly Journal of Experimental Psychology*. **64** (4): 807–822.

Risto Miikkulainen and Michael Dyer. 1991. Natural Language Processing with Modular PDP Networks and Distributed Lexicon. Cognitive Science, 15: 343-399.

Thomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 38(8):1388–1439. DOI: 10.1111/j.1551-6709.2010.01106.x

Gregory Murphy. 1988. Comprehending complex concepts. *Cognitive Science*, 12: 529-562.

Diane Pecher, Rene Zeelenberg, and Lawrence Barsalou. 2004. Sensorimotor simulations underlie conceptual representations Modality-specific effects of prior activation. *Psychonomic Bulletin & Review*, 11: 164-167.

David E. Rumelhart, James L. McClelland, and PDP Research Group (1986) *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, Volume 1: Foundations. Cambridge, MA: MIT Press.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. *In*

International conference on intelligent text processing and computational linguistics, 1-15. Springer, Berlin, Heidelberg.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 721-732.

Edward J. Wisniewski. 1997. When concepts combine. Psychonomic Bulletin & Review, 4, 167–183.

Edward J. Wisniewski. 1998. Property Instantiation in Conceptual Combination. *Memory & Cognition*, 26, 1330-1347.

Eiling Yee, & Sharon L. Thompson-Schill. 2016. Putting concepts into context. *Psychonomic Bulletin & Review*, 23, 1015–1027.