In Proceedings of the 6th Workshop on Cognitive Aspects of the Lexicon (CogALex-VI), Barcelona, ES, December 2020.

Characterizing Dynamic Word Meaning Representations in the Brain

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 USA
{naguirre,risto}@cs.utexas.edu

Abstract

During sentence comprehension, humans adjust word meanings according to the combination of the concepts that occur in the sentence. This paper presents a neural network model called CEREBRA (Context-dEpendent meaning REpresentation in the BRAin) that demonstrates this process based on fMRI sentence patterns and the Concept Attribute Representation (CAR) theory. In several experiments, CEREBRA is used to quantify conceptual combination effect and demonstrate that it matters to humans. Such context-based representations could be used in future natural language processing systems allowing them to mirror human performance more accurately.

1 Introduction

A word meaning is more than an entry in a dictionary. It involves a vast amount of knowledge relating the scenes and experiences people encounter (i.e., a rich encyclopedic knowledge), a set of referents to which the word properly applies (i.e., *the boy was angry* vs. *the chair was angry*), combination of other words, and grammatical constructions in which the word occurs. The meaning of the word varies from situation to situation and across contexts of use. For example, the word *small* means something different when used to describe a mosquito, a whale, or a planet. The properties associated with *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 (Medin & Shoben, 1988).

While humans have a remarkable ability to form new word meanings by combining existing concepts, modeling this process is challenging (Hampton, 1997; Janetzko 2001; Middleton et al, 2011; Murphy, 1988; Sag et al., 2002). The same concept can be combined to produce different meanings: *corn oil* means oil made of corn, *baby oil* means oil rubbed on babies, and *lamp oil* means oil for lighting lamps (Wisniewski, 1997, 1998). Since *lamp* is an object, oil is likely to be a member of the inanimate category. However, *corn* and *baby* are living things, which suggest otherwise. How do language users determine the membership structure of such combinations of concepts, and how do they deduce the interpretation? As this example illustrates, there is no simple rule on how to combine concepts (Cohen et al., 1984).

Computational models of such phenomena could potentially shed light into human cognition and advance AI applications that interact with humans via natural language. Such applications need to be able to understand and to form by themselves novel combinations of concepts. Consider for example virtual assistants such as Siri, OK Google, or Alexa. These applications are built to answer questions posed by humans in natural language. All of them have natural language processing software to recognize speech and to give a response. However, whereas humans process language at many levels, machines process linguistic data with no inherent meaning. Given the ambiguity and flexibility of human language, modeling human conceptual representations is essential in building AI systems that interact effectively with humans.

Today's experimental methods allow studying neural mechanisms underlying the semantic memory system. Neuroimaging (fMRI) technology, for instance, provides a way to measure brain activity during

word and sentence comprehension. When humans listen or read sentences they use different brain systems to simulate seeing the scenes and performing the actions that are described. As a result, parts of the brain that control these actions light up in the fMRI. Hence, semantic models have become a popular tool for prediction and interpretation of brain activity.

Recently, Machine Learning systems in vision and language processing have been proposed based on single-word vector spaces (Mikolov et al., 2013; Vinyals et al., 2015). They are able to extract low-level features in order to recognize concepts (e.g. cat), but such representations are shallow and fall short from symbol grounding (meaning). In general, these models build semantic representations from text corpora, where words that appear in the same context are likely to have similar meanings (Baroni et. al., 2010; Burgess, 1998; Devlin et al., 2018; Harris, 1970; Landauer & Dumais, 1997; Mikolov et al., 2013 Peters et al., 2018;). This problem has driven researchers to develop new componential approaches where concepts are represented by a set of basic features, integrating different modalities like textual and visual inputs. (Anderson et al., 2019; Bruni et al., 2012; Silberer & Lapata, 2014, Vinyals et. al., 2015). However, even with these multimodal embedding spaces, such vector representations lack intrinsic meaning, and therefore sometimes different concepts may appear similar.

A truly multimodal representations should account for the full array of human senses (Bruni et al., 2014). Embodiment theories of concept representation provide such an array (Barsalou, 1987; Binder et al., 2009; Landau et al., 1998; Regier, 1996). They allow for a direct analysis in terms of sensory, motor, spatial, temporal, affective, and social experience. Further, these theories can be mapped to brain systems. Recent fMRI studies helped identify a distributed large-scale brain network of multimodal sensory systems linked to the storage and retrieval of conceptual knowledge (Binder et al., 2009). This network was then used as a basis for Concept Attribute Representation (CAR) theory (a.k.a. the experiential attribute representation model). This theory is a semantic approach that represents concepts as a set of features that are the basic components of meaning, and grounds them in brain systems (Binder et al., 2009, 2011, 2016a, 2016b).

An intriguing challenge to semantic modeling is that concepts are dynamic, i.e. word meaning depends on context and recent experiences (Barsalou et al., 1993; Pecher et al., 2004; Yee et al., 2016). For example, a pianist would invoke different aspects of the word *piano* depending on whether he will be playing in a concert or moving the *piano*. When thinking about a coming performance, the emphasis will be on the piano's function, including sound and fine hand movements. When moving the piano, the emphasis will be on shape, size, weight and other larger limb movements (Barclay et al., 1974).

This paper addresses the challenge of dynamic representations based on CAR theory. The assumption is that words in different sentences have different representations. Therefore, different features in CARs should be weighted differently depending on context, that is, according to the combination of concepts that occur in the sentence. A neural network model is used to map brain-based semantic representations of words (CARs) into fMRI data of subjects reading everyday sentences. The goal is to identify how the weightings of the attributes in the CARs change to account for context (Aguirre-Celis & Miikkulainen, 2017, 2018, 2019, 2020). In this paper, the CAR theory is first reviewed, and the sentence collection, fMRI data, and word representation data described. Then, the computational model is presented followed by three evaluation studies: an individual example on the conceptual combination effect on word meanings, an aggregate study across the entire corpus of sentences, and a behavioral analysis to evaluate the neural network model.

2 Modeling Framework

To understand how word meanings change under the context of a sentence, three issues are addressed: (1) How are concepts represented? Componential theories of lexical semantics assume that concepts consist of a set of features that constitute the basic components of meaning. CAR theory represents such features in terms of known brain systems, relating semantic content to systematic modulation in neuroimaging activity. (2) How do word meanings change in the context of a sentence? A word is broken into various features that can become active at different rates in different situations. According to CAR theory, the weights given to different feature dimensions are modulated by context. (3) What tools and approaches can be used to quantify such changes? CAR theory assumes that context modifies the

baseline meaning of a concept. A computational model can test this assumption by using sentence fMRI patterns and the CAR semantic feature model to characterize how word meanings are modulated within the context of a sentence. The first two issues are addressed by the CAR theory. The third issue is addressed by CEREBRA, or Context-dependent mEaning REpresentation in the BRAin, a neural network model based on CAR theory.

2.1 Concept Attribute Representation (CAR) Theory

CAR theory is a semantic approach that represents concepts as a set of features that are the basic components of meaning (Anderson et al 2016, Binder, 2016a; Smith et al, 1974). 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. The features directly relate semantic content to systematic modulation of neuroimaging activity. This theory has been mostly applied to the task of prediction of neural activity patterns for individual concepts and entire sentences (Anderson et al., 2016, 2017, 2018, 2019; Binder et al., 2009, 2011, 2016a, 2016b, Fernandino et al., 2015).

Each word is modeled as a collection of 66 features that captures the strength of association between each neural attribute and word meaning. Furthermore, 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. Thus, people weigh concept features differently to construct a representation specific to the combination of concepts in the sentence.

Figure 1 shows the weighted CARs for the generic representation of the concept *bicycle*. The weight values represent average human ratings for each feature. For a more detailed account of this theory see Binder et al. (2009, 2011, 2016a, and 2016b).

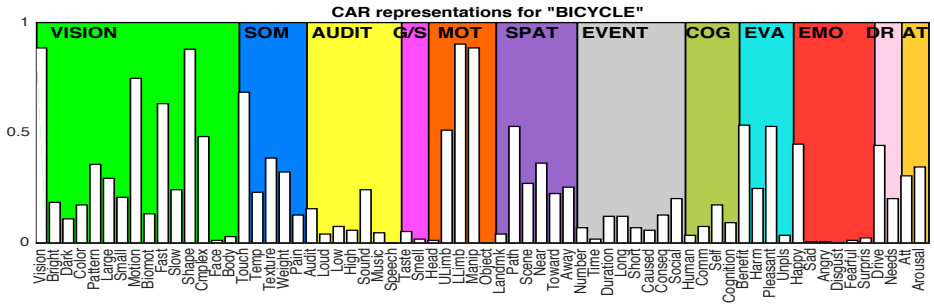

Figure 1: Bar plot of the 66 semantic features for the word *bicycle* (Binder et al., 2009, 2011, 2016a,2016b). It has low weightings on animate attributes such as Face, Body, and Speech, and emotions including Sad, and Fear and high weighting on attributes like Vision, Shape, Touch, and Manipulation. Similarly, it includes high weightings in Motion, Fast, Lower Limb and Path, since *bicycle* is considered a vehicle. CARs for *bicycle*.

2.2 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, the last two were collected by the Medical College of Wisconsin (Anderson et al., 2016; Binder et al., 2016a, 2016b). Additionally, fMRI representations for individual words (called SynthWord) were synthesized by averaging the sentence fMRI.

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*, *boy*, *hospital*, *desk*, *red*, *flood*, *damaged*, *drank*, *agreement*, *happy*, *hurricane*, *summer*, *chicken*, and *family*. An example of a sentence containing some of those words is *The flood damaged the hospital*.

Semantic Word Vectors: The 242 words (CAR) ratings were collected through Amazon Mechanical Turk (Anderson et al., 2016; Binder et al., 2016a). 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 *bicycle* 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 (Figure 1). In this manner, the representations map the conceptual content of a word to the corresponding neural representations, unlike other systems where the features are extracted from text corpora and the meaning is determined by associations between words and between words and contexts (Burgess, 1998; Landauer & Dumais, 1997; Mikolov et al., 2013).

Neural fMRI Sentence Representations: To obtain the neural correlates of the 240 sentences, subjects viewed each sentence on a computer screen while in the fMRI scanner. 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.

Eleven subjects took part in this experiment producing 12 repetitions each. 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). Due to noise inherent in the neural data, only eight subject fMRI patterns were used for this study. To form the target for the neural network, the most significant 396 voxels per sentence were then chosen (to match six case-role slots of the content words consisting of 66 attributes each) and scaled to [0.2..0.8].

Synthetic fMRI Word Representations: The neural data set did not include fMRI images for words in isolation. 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, 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. 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).

3 Computational Model

CEREBRA model was developed to investigate how words change under the context of a sentence using imaging data (Figure 2). It is based on the CAR semantic feature model and the FGREP neural network architecture (Forming Global Representations with Extended Backpropagation; Miikkulainen & Dyer, 1991). The model is trained to predict fMRI patterns of subjects reading everyday sentences. The FGREP mechanism is used to determine how the CARs would have to change to predict the fMRI patterns more accurately. These changes represent the effect of context; it is thus possible to track the brain dynamic meanings of words by tracking how the CARs feature-weightings change across contexts.

More specifically, the model is first trained to map CARWords (word attribute ratings) to SynthWords (fMRI synthetic words). Once it has learned this task, it is used to modify CAR words in context. SynthWords are combined to form SynthSent for the predicted sentence by averaging all words in the sentence. The SynthSent is then compared to the actual fMRISent (original fMRI data), to form a new error signal. That is, for each sentence, the CARWords are propagated and the error is formed as before, but during backpropagation, the network is no longer changed. Instead, the error is used to change the CARWords themselves (which is the FGREP method; Miikkulainen & Dyer 1991). This modification can be carried out until the error goes to zero, or no additional change is possible (because the CAR attributes are already at their max or min limits). Eventually, the revised CARWord represents the word meaning in the current sentence.

The CEREBRA model was trained 20 times for each of the eight fMRI subjects with different random seeds. A total of 20 different sets of 786 context word representations (one word representation for each sentence where the word appears) were thus produced for each subject. Afterwards, the mean of the 20 representations was used as the final representation for each word (per subject). It is important to emphasize that the goal of the CEREBRA model is not to predict the fMRI patterns as accurately and generally as possible, instead, it is used as a framework to identify and measure context-dependent changes in the CAR words (Aguirre-Celis & Miikkulainen, 2017, 2018, 2019, 2020).

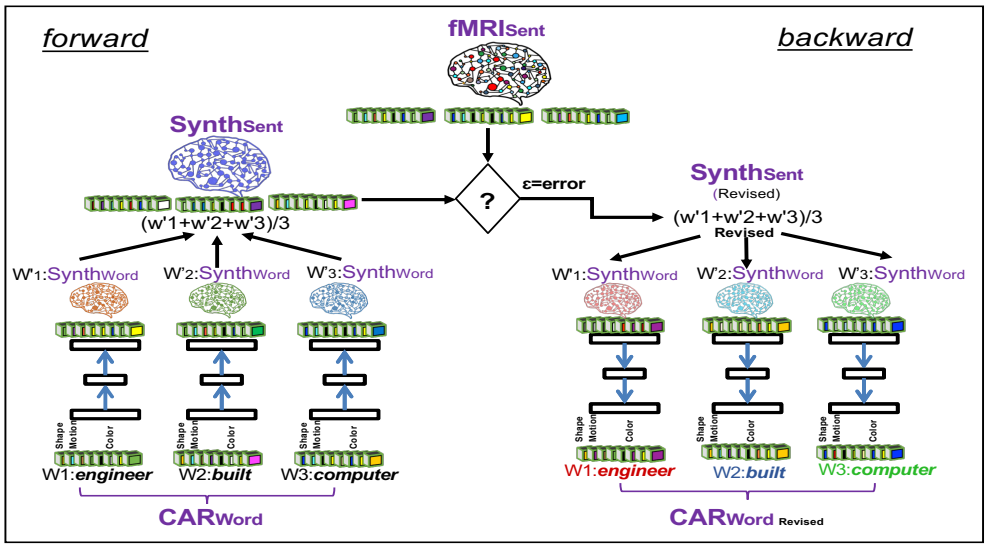

Figure 2: The CEREBRA model to account for context effects. After the model has been trained to map CARWords to SynthWords, it is used to determine how CAR words change in context. (1) Propagate CARWords to SynthWords. (2) Construct SynthSent by averaging the SynthWords into a prediction of the sentence. (3) Compare SynthSent with the observed sentence 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. Thus, the CEREBRA model captures context effects by mapping brain-based semantic representations to fMRI sentence images.

4 Experiments and Results

To evaluate the performance of CEREBRA as well as the context-based representations, two computational experiments and a behavioral analysis were conducted. The first two experiments measure how the CAR representation of a word changes in different sentences, and correlates 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. Next, an individual example of the conceptual combination effect is first presented, followed by the aggregate analysis and the behavioral study.

4.1 Analysis of an Individual Example

In the CAR theory, concepts' interaction arises within multiple brain networks, activating similar brain zones for both concepts. These interactions determine the meaning of the concept combination (Binder, 2016a, 2016b). As an example, consider the noun-verb interactions in Sentence 200: *The yellow bird flew over the field*, and Sentence 207: *The red plane flew through the cloud*. Since *bird* is a living thing, animate dimensions related to agency such as sensory, gustative, motor, affective, and cognitive experiences are expected to be activated, including attributes like Speech, Taste, and Smell. In contrast, *plane flew* is expected to activate inanimate dimensions related to perceiving an object, as well as Emotion, Cognition, and Attention.

Figure 3 shows the CARs for the word *flew* in the two sentences after they were modified by CEREBRA as described in Figure 2 and averaged across all eight subjects. In Sentence 200 there were indeed high activations on animate attributes like Pain, Smell and Taste, Audition, Music, Speech, as well as Communication and Cognition. In contrast, Sentence 207 emphasizes perceptual features like Color, Size, and Shape, Weight, Audition, Loud, Duration, Social, Benefit, and Attention.

The effect of conceptual combination on word meaning is clearly seen in this example. As the context varies, the overlap on neural representations create a mutual enhancement, producing a difference between animate and inanimate contexts. The CEREBRA model encodes this effect into the CAR representations where it can be measured. In other experiments, a similar effect was observed for other noun-verb pairs, as well as several adjective-noun pairs. Next, this effect is quantified statistically across the entire corpus of sentences.

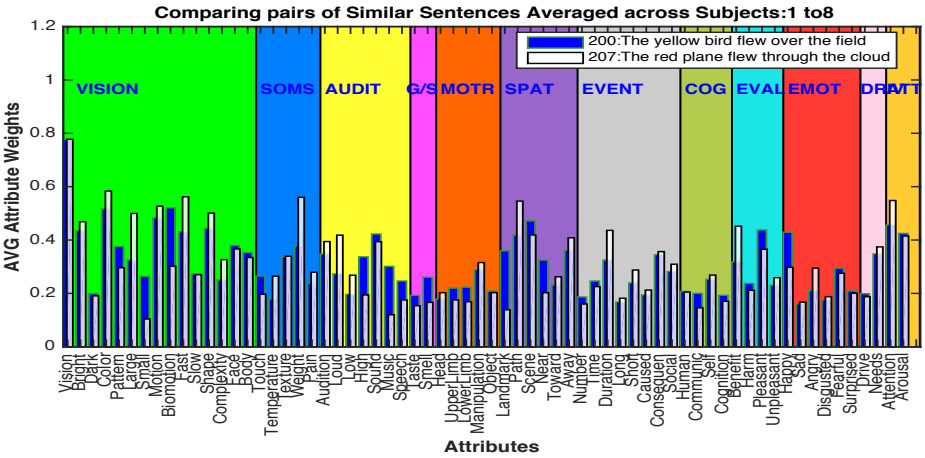

Figure 3: Contrasting the conceptual combination effect in two different sentences. In Sentence 200 (blue bars), the CAR representation modified by FGREP for the word *flew* has salient activations on animate features, likely denoting *bird* properties like Pain, Smell and Taste, and Communication. In Sentence 207 (white bars), it has high activations on inanimate object features, describing a Loud, Large, and Heavy object such as a *plane*. Thus, there is a clear difference between animate and inanimate features found in each sentence.

4.2 Aggregation Analysis

The aggregation study hypothesis is based on the idea that similar sentences have a similar effect, and this effect is consistent across all words in the sentence. This effect was verified in the following process:

1. For each subject, modified CARs for each word in each sentence were formed through FGREP as described in Figure 2.
2. A representation for each sentence, SynthSent, was assembled by averaging the modified CARs.
3. Agglomerative hierarchical clusters of sentences were formed using the set of SynthSents. The Ward method and Euclidean metric were used to measure the distance between clusters and observations respectively. The process was stopped at 30 clusters, i.e., at the point where the granularity appeared most meaningful (e.g., sentences describing open locations vs. closed locations).
4. Each cluster of sentences is expected to reveal similar changes in some of the dimensions. To recognize such common patterns of changes, the next step is to calculate the average of the changes for words with similar roles, e.g., *hospital*, *hotel*, and *embassy* (within the same cluster of sentences). That is, measure the difference between the new and the original CAR representations for each similar word roles and perform a statistical significance using *t*-test across the entire set for each CAR dimensions.
5. The modified CARs of the OWS were averaged.
6. Pearson's correlations were then calculated between the modified CARs and the average CARs of the OWS across all the dimensions.
7. Similarly, correlations were calculated for the original CARs.
8. These two correlations were then compared. If the modified CARs correlate with the CARs of the OWS better than the original CARs, context effect based on conceptual combination is supported.

In other words, this process aims to demonstrate that changes in a target word CAR originate from the OWS. For example, if the OWS have high values in the CAR dimension for Music, then that dimension in the modified CAR should be higher than in the original CAR for such target word. The correlation analysis measures this effect across the entire CAR representations. It measures whether the word meaning changes towards the context meaning. For more detail see (Aguirre-Celis & Miikkulainen, 2019).

The results are shown in Figure 4. 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.

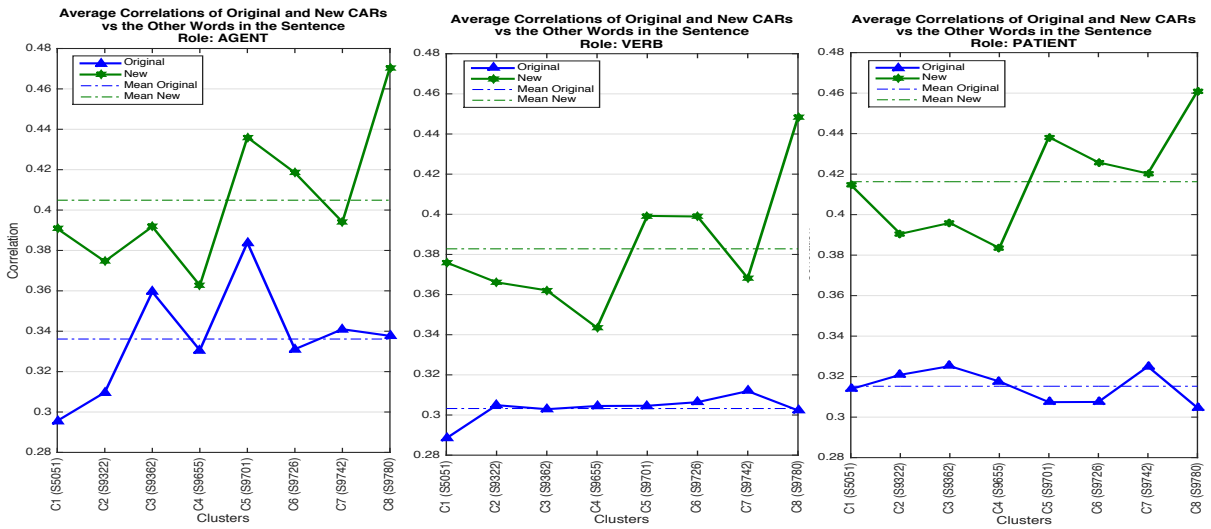

Figure 4: Correlation results. Average correlations analyzed by word class for eight subjects comparing original and new CARs vs. the average of the OWS respectively. A moderate to strong positive correlation was found between new CARs and the OWS, suggesting that features of one word are transferred 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.

4.3 Mapping Brain to Behavior

A behavioral analysis was designed to evaluate the CEREBRA's context-based representations via human judgements. That is, Sections 4.1 and 4.2 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 are compared to changes predicted by the CEREBRA model.

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 in accordance with the University of Texas at Austin Institutional Review Board (2018-08-0114).

The complete survey is an array of 24 questionnaires that include 15 sentences each. For each sentence, the survey measures 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 contains 150 evaluations. For example, a questionnaire might measure changes on 10 specific attributes such as 'is visible', 'living thing that moves', 'is identified by sound', 'has a distinctive taste', for a specific word class such as *politician*, for 15 sentences such as *The politician celebrated at the hotel*. A particular example sentence questionnaire is shown in Figure 5.

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 the table 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 Model 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 related to conceptual combination (Hampton, 1997; Wisniewsky, 1998), contextual modulation (Barclay, 1974), and attribute centrality (Medin & Shoben, 1988): the same property is true for two different concepts but more central to one than to the other (e.g., it is more important for boomerangs to be curved than for bananas).

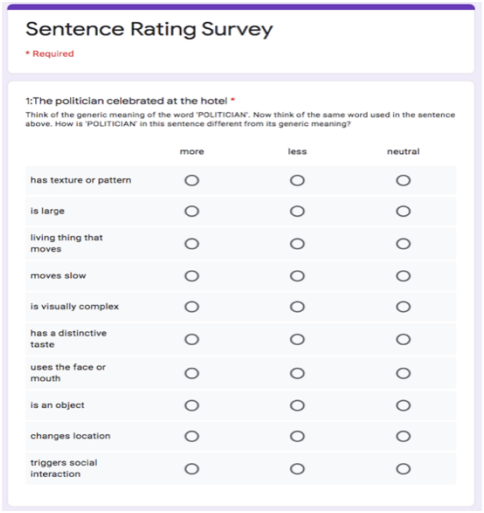

Figure 5: An example sentence in the survey. The sentence is *The politician celebrated at the hotel*, the target word is *politician* in the role of Agent. Ten different attribute changes are measured by selecting whether the attribute increased ("more"), decreased ("less") or remained "neutral". These human judgements were then matched with those predicted by CEREBRA.

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
-1/0	1074	466	8
1	892	587	886
TOTAL	1966	1052	894
Match each other		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.

The direction of change is therefore not a good predictor of human responses. Instead these changes need to be measured relative to changes in the OWS. Three approaches were thus used to evaluate the changes: (1) What is the effect of the rest of the sentence in the target word? This effect was measured by computing the average of the CEREBRA changes (i.e., new-original) of the OWS, and subtracting that average change from the change of the target word. (2) What is the effect of the entire sentence in the target word? This effect was measured by computing the average of the CEREBRA changes (i.e., new-original) of all the words in the sentence including the target word, and subtracting that average change from the change of the target word. (3) 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-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 Model 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 three CEREBRA approaches produced very similar results, therefore only those of the third approach 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 statistically strongly significant for all of 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, 2020).

5 Discussion and Future Work

An interesting future work direction would be to replicate the study on a more extensive data set with a fully balanced stimuli and with fMRI images of individual words. The differences should be even stronger and it should be possible to uncover more refined effects. Such data should also improve the survey, since it would be possible to identify questions where the effects can be expected to be more reliable.

Compared to other approaches, such as distributional semantic models (DSMs), CAR theory enables a mapping between conceptual content and neural representations. In CARs conceptual knowledge is distributed across a small set of modality-specific neural systems that are engaged when instances of the concept are experienced. In contrast, DSMs reflect conceptual knowledge acquired through a lifetime of linguistic experience, and they are not grounded on perception and action. Experiential data specify the perceived physical attributes or properties associated with the referents of words (e.g., a carrot refers to an object whose attributes describes it as orange, conical/cylindrical, juicy, crispy, sweet). In contrast, linguistic data specify how a given word is statistically distributed across different texts (e.g., a carrot is a root vegetable, usually orange, Dutch invented the orange carrots, it contains high carotene, human body turns carotene into vitamin A). A lot of experiential data is usually unstated in such texts. Thus, experiential data provide a foundation that support both perceptual data (e.g., answering "orange" to "What color are carrots?), as well as associative/encyclopedic data (e.g., answering "rabbit" to "What animal likes to eat carrots?"; Anderson et al., 2019; Andrews et al., 2009; Martin, 2007).

In the future, multimodal CEREBRA representations could be used to make natural language processing systems more robust. For instance, it may be possible to train a neural network to represent context simultaneously from both DSMs and CEREBRA representations as part of a natural language understanding system for service robot applications. For instance, service robots with such representations would have the capability to understand natural language commands (e.g., watering plants), to have encyclopedic knowledge (i.e., to make decisions), to ground language by adapting to the environment (i.e., object recognition, location) and by understanding novel concepts (i.e., "rain water"). Thus, the CEREBRA representations provide the experiential-based data (i.e., concrete words) and the DSMs provide the association-based data (i.e., abstract words), leading to a more robust performance.

6 Conclusion

The CEREBRA model was constructed to test the hypothesis that word meanings change dynamically based on context. The results suggest three significant findings: (1) context-dependent meaning representations are embedded in the fMRI sentences, (2) they can be characterized using brain-based semantic representations (CARs) together with the CEREBRA model, and (3) the attribute weighting changes are real and meaningful to the subjects. CEREBRA thus takes a step towards understanding how the brain constructs sentence-level meanings dynamically from word-level features.

Acknowledgements

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. (2020). Characterizing the Effect of Sentence Context on Word Meanings: Mapping Brain to Behavior. *Computation and Language*. arXiv:2007.13840.

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*.

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

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.

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.

Benjamin Cohen & Gregory Murphy. (1984). Models of Concepts. *Cognitive Science 8:*25-78.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computation and Language*. arXiv:1810.04805

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.

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.

Barbara Landau, Linda Smith, and Susan Jones. 1998. Object Perception and Object Naming in Early Development. Trends in Cognitive Science, 27: 19-24.

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

Alex Martin. 2007. The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25–45.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. 2018. Deep contextualized word representations. *Computation and Language*. arXiv:1802.05365.

Terry Regier. 1996. The Human Semantic potential. *MIT Press*, Cambridge, MA.

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 E. Smith, Edward J. Shoben, and Lance J. Rips. 1974 Structure and process in semantic memory: A featural model for semantic decisions. *Psychological review* 81:214.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A New Image Caption Generator. *Computing Research Repository*, arXiv:1506.03134v2

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-Schil. 2016. Putting concepts into context. *Psychonomic Bulletin & Review*, 23, 1015–1027.