From Words to Sentences & Back: Characterizing Context-dependent Meaning Representations in the Brain

Nora Aguirre-Celis (naguirre@cs.utexas.edu)
ITESM, Monterrey, Mexico &
Department of Computer Science, 2317 Speedway
Austin, TX 78712 USA

Manuel Valenzuela-Rendon (valenzuela@itesm.mx)
ITESM, Monterrey, Mexico

Risto Miikkulainen (risto@ cs.utexas.edu)
Department of Computer Science, 2317 Speedway
Austin, TX 78712 USA

Abstract

Recent Machine Learning systems in vision and language processing have drawn attention to single-word vector spaces, where concepts are represented by a set of basic features or attributes based on textual and perceptual input. However, such representations are still shallow and fall short from symbol grounding. In contrast, Grounded Cognition theories such as CAR (Concept Attribute Representation; Binder et al., 2009) provide an intrinsic analysis of word meaning in terms of sensory, motor, spatial, temporal, affective and social features, as well as a mapping to corresponding brain networks. Building on this theory, this research aims to understand an intriguing effect of grounding, i.e. how word meaning changes depending on context. CAR representations of words are mapped to fMRI images of subjects reading different sentences, and the contributions of each word determined through Multiple Linear Regression and the FGREP nonlinear neural network. As a result, the FGREP model in particular identifies significant changes on the CARs for the same word used in different sentences, thus supporting the hypothesis that context adapts the meaning of words in the brain. In future work, such context-modified word vectors could be used as representations for a natural language processing system, making it more effective and robust.

Keywords: Neural Networks; FGREP; Concept Attribute Representation theory; fMRI; Context; Meaning; Semantics; Embodied Cognition

Introduction

Recently, Deep Learning systems of vision and natural language processing (NLP) have drawn special attention into single-word vector spaces. They are able to extract low level features in order to recognize concepts (e.g. cat), but they are incapable of forming an abstract notion of the concept (symbol). In general, these models build semantic representations from text corpora where words that appear in the same context are likely to have similar meanings (Harris, 1970; Landauer & Dumais, 1997, Burgess, 1998; Baroni et. al., 2010). However, such representations lack intrinsic meaning, which means sometimes even different concepts may appear similar. This problem has driven researchers to develop new componential approaches, where concepts are represented by a set of basic features, or attributes, based on textual and perceptual input. (Bruni, et al., 2012; Silberer & Lapata, 2014, Vinyals et. al., 2015). However, even with their multimodal embedding space, such vector representations fall short from symbol grounding.

In contrast, embodiment theories of knowledge representation (Regier, 1996; Landau et al., 1998, Barsalou, 2008) provide a direct analysis in terms of sensory, motor, spatial, temporal, affective, and social phenomena. Further, these theories can be mapped to brain networks. Recent fMRI studies helped identify a distributed large-scale network of sensory association, multimodal and cognitive regulator systems linked with the storage and retrieval of conceptual information (Binder et al., 2009). This network was then used as a basis for Concept Attribute Representation (CAR) theory, an embodiment theory that enumerates semantic features of concepts and grounds them in brain networks (Binder et al., 2009, 2011 and 2016).

An intriguing challenge to such theories is that concepts are dynamic, i.e. word meaning depends on context and recent experience (Pecher, Zeelenberg, & Barsalou, 2004). 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.

This paper focuses on addressing these challenges based on the CAR theory. The main idea is that different attributes in CARs can be weighted differently depending on context, i.e. according to how important each attribute is in that context. More specifically, neutral CARs of words are first used to form an expected fMRI pattern of a subject reading a sentence. That pattern is compared to an actual fMRI image. Two techniques, multiple linear regression and a FGREP neural network, are then used to determine how the CARs would have to change to account for the actual fMRI

pattern. These changes represent the weighting in context; it is thus possible to track the dynamic meanings of words by tracking how the weighting changes across contexts.

Experiments with available fMRI data show that the approach is feasible, demonstrating meaningful differences for e.g. *human communication* vs. *noise from a machine*; *dangerous storm* vs. *dangerous person*; *live mouse* vs. *dead mouse*. These changes are principled and could be captured e.g. by a neural network. It might then be possible to create them dynamically, and form as a basis for a more robust and grounded natural language processing system.

The CAR theory is first reviewed below, and the sentence fMRI and word representation data described. The methods for determining semantic changes, i.e. multiple linear regression and FGREP, are then presented, followed by an analysis of the results.

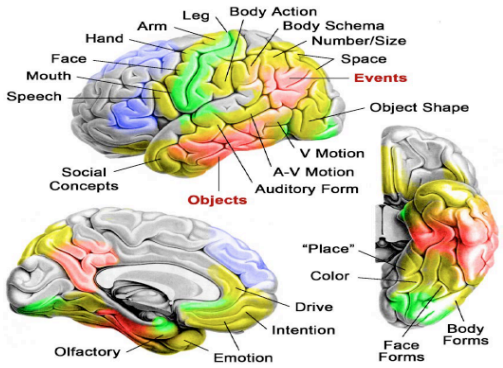

Figure 1: Perceptual Grounding. CARs are composed of a list of known modalities that relate to specialized sensory, motor and affective brain processes, systems processing spatial, temporal, and casual information, and areas involved in social cognition. They capture aspects of experience central to the acquisition of abstract and concrete event as well as object concepts.

Concept Attribute Representation Theory

CARs represent the basic components of meaning defined in terms of known neural processes and brain systems (Binder, 2016). They relate semantic content to systematic modulation in neuroimaging activity. And are therefore not limited to the classical sensory-motor dimensions of most embodied theories.

CARs 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. They capture aspects of experience central to the acquisition of event and object concepts (both abstract and concrete).

These attributes were selected after an extensive body of physiological evidence based on two assumptions: (1) all aspects of mental experience can contribute to concept acquisition and consequently concept composition; (2) experiential phenomena are grounded on neural processors representing a particular aspect of experience (Figure 1).

These aspects of mental experience model each word as a collection of a 66-dimensional feature vector that captures the strength of association between each neural attribute and the word meaning. An example is shown in Figure 2. For a more detailed account of the attribute selection and definition see Binder et al. (2009, 2011 and 2016).

Data Collection and Preprocessing

Two existing data sets were used in this study: fMRI images of sentences and CARs obtained via Mechanical Turk.

Neural Images

The stimuli shown to subjects consisted of a list of 240 every day written sentences prepared in the Knowledge Representation in Neural Systems (KRNS) project (Glasgow et al., 2016). The sentences are composed by three to nine words from a set of 242 words (141 nouns, 39 adjectives and 62 verbs). Eleven subjects took part in this experiment producing 12 repetitions each. Participants viewed the sentences word by word while in the scanner. The data was acquired by the Center for Imagining Research of the Medical College of Wisconsin (Anderson et al., 2016). The fMRI data was preprocessed and transformed into a single sentence fMRI representation per participant (by averaging all the repetitions), with a final selection of 396 voxels per sentence on a scale from 0.2-0.8, for further use in the computational models.

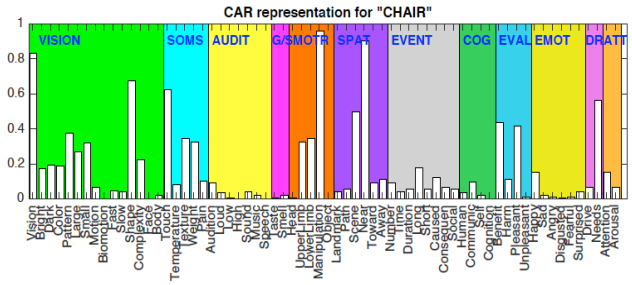

Figure 2: Bar plot for CAR 66 semantic features. The attribute ratings represent the basic features of *chair*. Given that this concept is an object, gets low weightings on human-related attributes: face, speech, social, and emotion and strong on visual, shape, touch, manipulation, and some others.

Semantic Vectors

The semantic attribute ratings were collected thru Amazon Mechanical Turk for each of the 242 words (e.g. *family, hospital, chair, small, green, laughed, listened, walked*). In a scale of 0..6, the participants were asked to assign the degree to which a given concept is associated to a specific type of neural component of experience (e.g. "To what degree do you think of a *chair* as having a fixed location, as on a map?"). Approximately 30 ratings (all attributes for each word) were collected. After averaging all the ratings and removing outliers, the final attributes were transformed to unit length yielding a collection of 66-dimensional feature vector that captures the weights of association between each neural attribute and the 242 words. Note that in this manner, the richness and complexity of representations is based on intrinsic meaning of each word, and not on word co-occurrence (Figure 2).

Data Preparation

The data set did not include fMRI images for words in isolation, a technique developed by Anderson et al. (2016) was adopted to approximate them. The voxel values for words were obtained by averaging all fMRI images for the sentence where each word occurred. Thus, the vectors include a combination of examples of that word along with other words that appear in the same sentence (context). Because of the limited number of combinations, some of these became identical, and were excluded from the dataset.

Given the final set of 237 sentences and 236 words (138 nouns, 38 adjectives and 60 verbs), the next step was to identify pairs of sentences with differences on word meanings such as *live mouse* vs. *dead mouse*, *good soldier* vs. *soldier fighting*, *built hospital* vs. *damaged hospital,* and *playing soccer* vs. *watching soccer*. This collection will allow the computational models to evaluate distinctive attribute representations and consequently adjust the base-line meaning of a word to convey the effects of context and conceptual combination.

A collection of 77 such sentences, with different shades of meaning for verbs, nouns and adjectives, as well as different contexts for nouns and adjectives was assembled. This collection will be used as Words of Interest (WoI) for the analysis of context in the experiments (Table 1).

Table 1: Contrasting Sentences. Eight sentences from the collection of the 77 contrasting sentences. Here, for instance, the verb *kicked* is used in two different contexts, playing with a ball (as in Soccer) vs. breaking the door (as an aggressive behavior).

SEMANTIC CONTRAST	SENTENCES
GOOD	94 *The soldier delivered the medicine during the flood.*
AGGRESSIVE	112 *The soldier kicked the door.*
PLAY (SOCCER)	239 *The artist kicked the football.* 62 *The boy kicked the stone along the street.*
BREAK	112 *The soldier kicked the door.*
BAD PEOPLE	119 *The dangerous criminal stole the television.* 152 *The mob was dangerous.*
NATURE	99 *The flood was dangerous.*

Computational Models

A new technique is proposed in this section for analyzing data imaging. It is grounded on the CAR theory and implemented using Multiple Linear Regression (LReg) and the FGREP neural network (Forming Global Representations with Extended BP; Miikkulainen & Dyer, 1991). The main idea is to predict sentence fMRI by mapping CARWord to SynthWord (fMRI) (top of Figure 3). The SynthWord is then combined by averaging to form SyntSent for the predicted sentence. Next, the SynthSent is compared to the actual fMRISent (middle of Figure 3). The differences are included by modifying the SynthWord that map to fMRISent and by modifying the CARWord that map to the modified SynthWord (bottom of Figure 3). The resulting CARWord indicate how word meaning change across sentences.

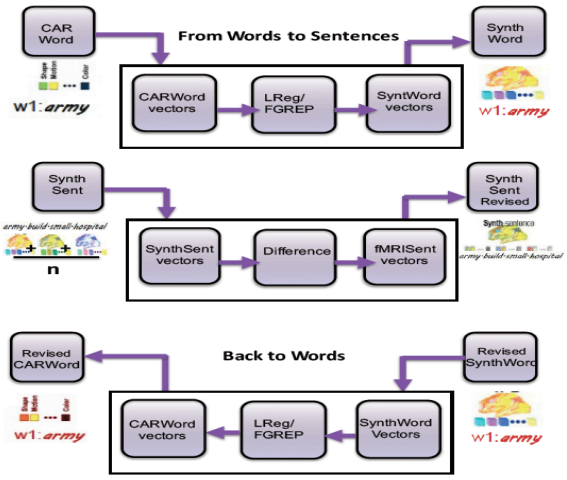

Figure 3: General System framework and data flow. Mapping CARWord to SynthWord (top). Then SynthWord is combined by averaging to form SyntSent and to be compared to the actual fMRISent (middle). Invert the process to modify the CARWords via SynthWord revised (bottom). The Revised CARWord includes different word meaning across sentences.

Multiple Linear Regression

At the word level, Multiple regression (LReg) is used to learn the mapping between CARWord and SynthWord voxels. The training set has attribute vectors of words as independent variables and the corresponding SynthWord vectors as the dependent variable, predicting one voxel at the time. Similarly, at the sentence level, the training contains assembled sentences (SynthSent) as independent and the corresponding Observed fMRISent as the dependent variable. Once the prediction error is calculated, LReg is inverted (which is possible because it is linear), to determine what the CARWord values should have been to make the error zero.

Neural Network with FGREP

It is possible that the linear prediction based on LReg is not powerful enough to account for the context effects. Therefore, a nonlinear approach based on neural networks is tested as well. A neural network is trained to map CARWord to SynthWord, which are then averaged (as before) into a prediction of the sentence SynthSent (Figure 4). The prediction error is used (through backpropagation) to train the network.

After training, this network is used to determine how the CARWords should change to eliminate the error. 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---Forming Global Representations through Extended backPropagation; Miikkulainen et al., 1991). This modification can be carried out until the error goes to zero, or no additional change is possible (because the CAR values are already at their max or min limits).

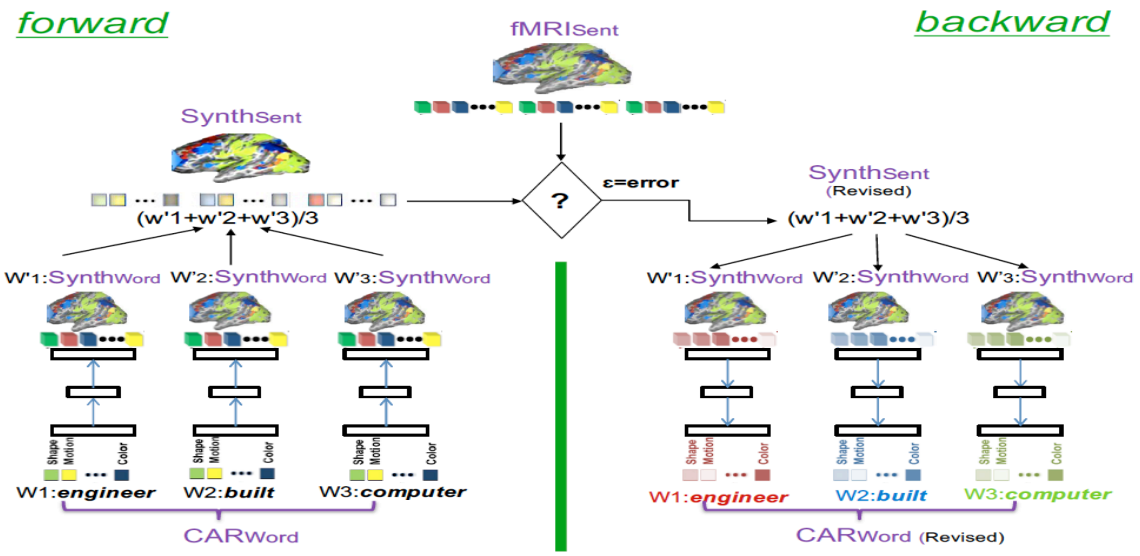

Figure 4: The FGREP model to account for context effects. Propagate CARWord to SynthWord. Compose SynthSent by averaging the words into a prediction of the sentence. Compare SynthSent against Observed fMRISent. Backpropagate the error with FGREP for each sentence, freezing network weights and changing only CARWord. Repeat until error reaches zero.

Training the neural network requires as input the 236 CARWord 66-dimensional vectors (W1, W2, W3) and as target, the equivalent corresponding 396-dimensional SynthWord vector (W'1, W'2, W'3). The network then learns a general mapping of words across all sentences. This mapping is then utilized in the FGREP phase to change the CARWord for each different sentence separately (Figure 4). As the last step, the changes in the semantic attributes are analyzed according to the CAR theory for each affected sentence. At this point, due to scarcity of data this is a manual process verifying that the changes made sense.

Results

The two approaches LReg and FGREP were evaluated in a preliminary experiment of distinguishing between the different meanings of the verb *listened*. LReg was found to be inadequate in this task and therefore in two subsequent experiments, focusing on the the adjective *dangerous* and in the noun *mouse* only the FGREP approach was used. The analysis was performed on the individual subjects for which the fMRI data in general was most consistent.

Different contexts for the verb "*listened* "

Both models were used in this experiment to compare the contrasting meanings of HUMAN COMMUNICATION vs. NOISE FROM A MACHINE for the word *listened* as expressed in 89: *The mayor listened to the voter*, 92: *The lonely patient listened to the loud television*. The left side of Figure 5 shows the results for LReg between the original and transformed CARs. Although the CARs adjusted in all sentences, the changes were small and unprincipled, unable to characterize the difference between human communication versus noise from a machine. In contrast, the outcome for FGREP resulted in context-dependent changes as shown, for sentences 89 and 92 in the right side of Figure 5.

CARs in Sentence 89 presented salient activations in human-related attributes like Face, and Body, Audition, and Speech, as well as Human, Communication, and Cognition, presumably denoting human verbal interaction. For Sentence 92, high activations on Vision, Bright, Color, Pattern, Large, Shape, Complexity, Touch, Temperature, Weight, Scene, Near, Harm, Unpleasant, Happy, and Angry describe a loud and large object such as a television. These results suggest that the linear mapping that LReg performs is not powerful enough to capture context, but the nonlinear mapping of FGREP is. The following experiments therefore both used the FGREP method for this task.

Different contexts for the adjective "*dangerous*"

This experiment compared the contrasting meanings of NATURE vs. BAD PEOPLE for the word "dangerous", as expressed in 98: *The flood was dangerous*, 118: *The dangerous criminal stole the television*. Figure 6 shows the differences resulting from the FGREP method. As with the verb *listened*, context-dependent changes did emerge.

CARs in Sentence 98 present changes on activation for Large, Motion, SOMS attributes Texture and Weight, and event attributes Time, Short, and Caused, reflecting moving water. The attributes Toward, Harm, Unpleasant, and the emotion of Angry, represent the experiential and personal nature of danger. Conversely, Sentence 118 shows high activation for Vision, Complexity, Face, and Speech, because they represent human types and roles such as a criminal. Motor attribute Lower Limb as well as evaluation attributes Benefit, Angry, Disgusted, and Fearful can be associated with a dangerous act by a criminal. The FGREP method, therefore, was largely able to differentiate between

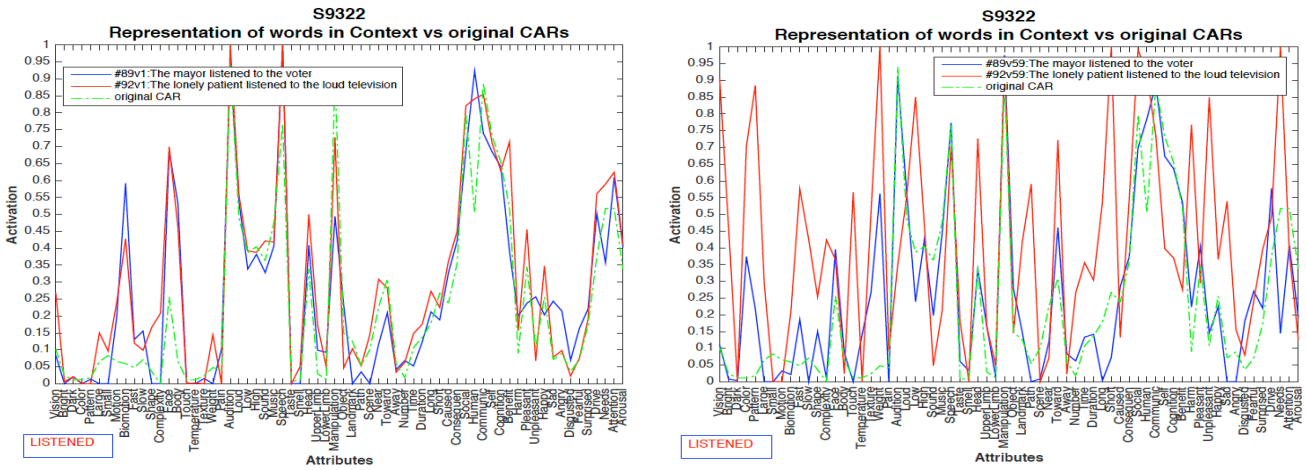

Figure 5: Results for the word *listened* in two contrasting sentences. LReg (left) did not capture context. All changes were insignificant to characterizing the context-dependent representations. The green line shows the original CARs for comparison. FGREP (right) did grasp context. The CARs for Sentence 89 have increased activations in human-related attributes like Face and Body, Auditory attributes, as well as Human, Communication and Cognition. In contrast, Sentence 92 activations on Vision, Color, Large, Shape, Complexity, Touch Temperature, High sound, and Unpleasant, depict a loud object such as a television.

the contrasting relevant dimensions of *dangerous* act of nature and humans.

Different contexts for the noun "*mouse*"

This experiment compared the contrasting meanings of DEAD vs. ALIVE for the word *mouse* as expressed in sentences 56: *The mouse ran into the forest*, 60: *The man saw the dead mouse*. Figure 7 shows the differences resulting from the FGREP method, which are again systematic and meaningful.

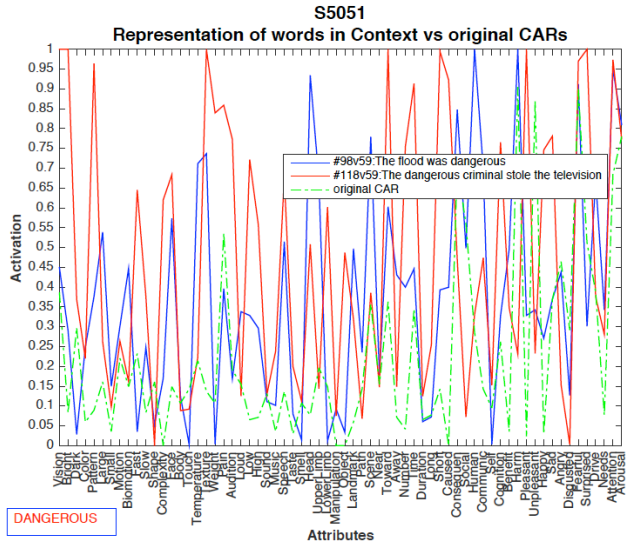

Figure 6: FGREP results for the adjective *dangerous* across two contrasting sentences. CARs in Sentence 98 changed activation for Large, Motion, Texture and Weight, Time, Short, and Caused, reflecting moving water. The attributes Toward, Harm, Unpleasant, and Angry, represent the experiential nature of danger. Sentence 118 shows high activation for Vision, Complexity, Face, and Speech, because they represent human types and roles. Lower Limb, Benefit, Angry, Disgusted and Fearful can be associated

CARs in Sentence 56 have increased activation for Vision, Motion, Complexity, High, and Sound, possibly suggesting animate properties of the live mouse. Upper Limb, spatial attributes Path and Away, and event attributes Time, Duration, Short, and Consequence, symbolize activity such as running. Emotions of Fearful and Surprised may well be associated with seeing a live mouse. In contrast, Sentence 60 shows increased activation for Temperature, Weight, and Smell, as well as emotions Sad, Angry, Disgusted and Fearful, which may be associated to the dead mouse. These changes indicate different aspects of mouse in two contrasting contexts.

Discussion and Further Work

The experiments in this paper suggest that different aspects of word meaning are activated in different contexts, and it is possible to see those changes in the corresponding fMRI images. These changes are likely to be nonlinear: The linear mapping approach (regression) tends to muddle them, but a nonlinear mapping (FGREP neural network) can tease them apart.

This result is remarkable considering that the dataset was not originally designed to answer the question of dynamic meaning. In particular, having fMRI images for isolated words available, instead of having to synthesize them, should amplify the observed effects significantly. It should also be possible to include sentences with contrasting contexts systematically, thus increasing the number of possible observations, and making it possible to identify differences in a more comprehensive manner.

With such a larger dataset, it should be possible to characterize changes across multiple sentences. Different kinds of changes may occur in nouns, adjectives, and verbs, and there are likely to be interactions between them. Moreover, the semantic changes can vary from individual to individual. As the first step, only single subjects were analyzed in this paper. In the future, the analysis can be

extended to more subjects, identifying which changes are consistent across subjects, and which ones are more individualistic. For instance, the subject in experiment 3 was Sad that the mouse was dead; another subject could show a different emotion.

After formulating such principles, the next step would be to utilize them in building artificial natural language processing systems. It may be possible to train e.g. a neural network to predict how meaning changes in context. Such a network could be then used as a part of an engineered natural language processing system, dynamically modifying the vector representations for the words to fit the context. Such a system should be more effective and more robust in its inference, and match human behavior better.

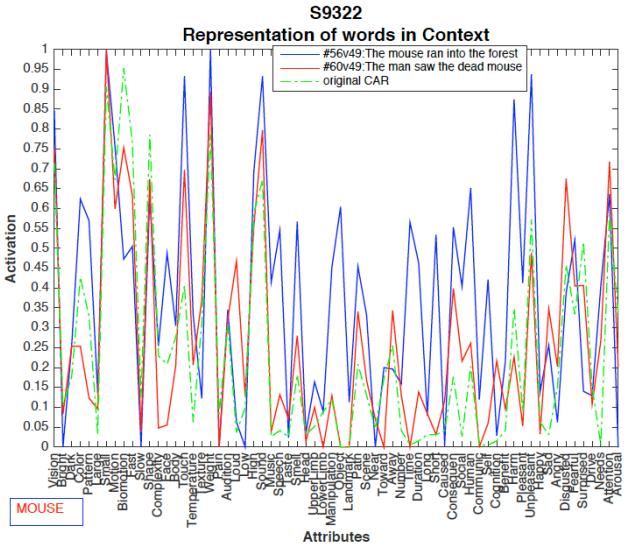

Figure 7: FGREP results for the noun *mouse* across two contrasting sentences. CARs in Sentence 56 increased activation for Vision, Motion, Complexity, High, and Sound, presumably to indicate the animate properties of the live mouse. Upper Limb, Path, Away, Time, Duration, Short, and Consequence, suggest activity such as running. In contrast, Sentence 60 shows increased activation for Temperature, Weight, and Smell, as well as Sad, Angry, Disgusted and Fearful, which can be associated to the dead mouse. These changes indicate different aspects of mouse in two contrasting contexts.

Conclusion

Concepts are dynamic; their meaning depends on context and recent experience. In this paper, word meaning was represented as a collection of attributes (CARs), grounded in observed brain networks. Multiple Linear Regression analysis and a nonlinear FGREP Neural Network were used to understand how the CARs could change to construct the actual sentence representations seen in fMRI images. Preliminary results suggest that there are indeed systematic changes in CARs, and they make sense in each sentence context. These changes could only be seen in the FGREP analysis, suggesting that they are likely to be nonlinear. In the future, such changes could be characterized more fully and used to make artificial natural language systems sensitive to context.

Acknowledgments

We would like to thank Jeffery Binder (Medical College of Wisconsin), Rajeev Raizada and Andrew Anderson (University of Rochester), Mario Aguilar and Patrick Connolly (Teledyne Scientific Company) for their work and valuable help regarding this research. This work was supported in part by IARPA-FA8650-14-C-7357 grant.

References

Anderson, A. J., Binder, J. R., Fernandino, L., Humpries C. J., Conant L. L., Aguilar M., Wang X., Doko, S., Raizada, R. D. (2016). Perdicting Neural activity patterns associated with sentences using neurobiologically motivated model of semantic representation. *Cer. Cortex*. 1-17. Doi:10.1093/cercor/bhw240.

Baroni, M., Murphi, B., Barbu, E., Poesio, M. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science,* 34(2):222-254.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology,* 59:617-845.

Binder, J. R., Desai, R. H., Graves, W. W., Conant L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19:2767-2769.

Binder, J. R., Desai, R. H. (2011). The neurobiology of semantic memory. *Trends Cognitive Sci*, 15(11):527-536.

Binder, J. R., Conant L. L., Humpries C. J., Fernandino L., Simons S., Aguilar M., Desai R. (2016). Toward a brain-based componential semantic representation. *Cog. Neuropsychology,* 33:3-4, 130-174.

Binder, J. R. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, 23.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with HAL model. *Behavior Research Methods, Inst. & Com.,* 30, 188-198.

Burni, E., Tran, N., Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. R. (JAIR),* 49:1-47

Glasgow, K., Roos, M., Haufler, A. J., Chevillet, M., A., Wolmetz, M. (2016). *Evaluating semantic models with word-sentence relatedness*. arXiv:1603.07253.

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

Landau, B., Smith, L., and Jones, S. (1998). Object Perception and Object Naming in Early Develop. *Trends in CosSci* 27:19-24.

Landauer, T.K., Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review,* 104, 211-240.

Miikkulainen, R., Dyer, M., G. (1991). Natural Language Processing with Modular PDP Networks and Distributed Lexicon. *Cognitive Science*, 15, 343-399.

Pecher, D., Zeelenberg, R., Barsalou, L. W. (2004). Sensorimotors simulations underlie conceptual representations: Modality-specific effects of prior activation. *Psychonomic Bulletin & Review, 11,* 164-167.

Regier, T. (1996). *The Human Semantic potential*. MIT Press, Cambridge, Massachusetts.

Silberer, C., Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 721-732.

Vinyals, O., Toshev, A., Bengio, S., Erham, D. (2015). Show and Tell: A New Image Caption Generator. arXiv:1506.03134v2.