Emergent Semantic Proto-role Structure in Tensor Product Representations

Devanshu Singh **Paul Smolensky**
Johns Hopkins University
{dsingh33,psmolen1}@jhu.edu

Abstract

The theory of semantic proto-roles, developed in Dowty (1991), replaces discrete semantic role representations with vectors composed of real-valued scalar annotations of proto-role properties, which are semantic entailments about the role played by an argument in an event described in a sentence. In this paper, we posit a deep conceptual link between the proto-role theory and the more general and comprehensive Integrated Connectionist/Symbolic Cognitive Architecture developed in Smolenksy and Legendre (2006). Specifically, this architecture theorizes Tensor Product Representations as the fundamental form of all cognitive representations. TPRs are explicitly compositional vector embeddings of symbolic structures and share key properties of both traditional discrete symbolic representations like parse trees and discrete semantic roles, as well as neural vector representations like word embeddings. Next, we test this theoretical link by training a TP-TRANSFORMER language model, which learns TPRs of sentences, and analyzing its learned parameters to find associations between specific vector embeddings and proto-role properties or clusters. In the spirit of recent work on using TPRs to make neural networks more explainable, we find that certain vectors used to compute TPRs within the model can be interpreted as corresponding to certain configurations of semantic proto-role properties.

1 Introduction

Semantic roles are entities, also called arguments and often syntactically evoked by nouns, involved in events, also called predicates and often syntactically evoked by verbs, that are described by many sentences. They often answer the question of "who did what to whom with what and where?" in reference to these events. For example,

Figure 1: Output from University of Pennsylvania Cognitive Computation Group's online demo of SRL model (Punyakanok et al. (2008)).

in the following sentence, which describes a hitting event, the arguments *Devanshu*, *baseball*, and *bat* play the semantic roles of AGENT, PATIENT, and INSTRUMENT respectively.

`Devanshu hit the baseball with his bat.`

Semantic roles are essential elements of an event-based theory of linguistic representation and important for the natural language processing (NLP) task of semantic role labeling (SRL), which traditionally focuses on classifying arguments into a set of discrete semantic roles. The semantic roles described above, AGENT, PATIENT, and INSTRUMENT, are three of the most common discrete semantic roles used in SRL. If the argument and/or event spans have been labelled and are part of the data given to an SRL model, then the NLP task reduces to a multi-class classification task or a sequence classification task and can be accomplished by many standard classification or tagging systems. If this data is not given, however, then SRL can become as complex as semantic parsing, in which the model itself must find all of the event and argument spans, their labels, and the structural relationships between them. In all cases, a discrete structure is predicted by a model from a discrete sequential input.

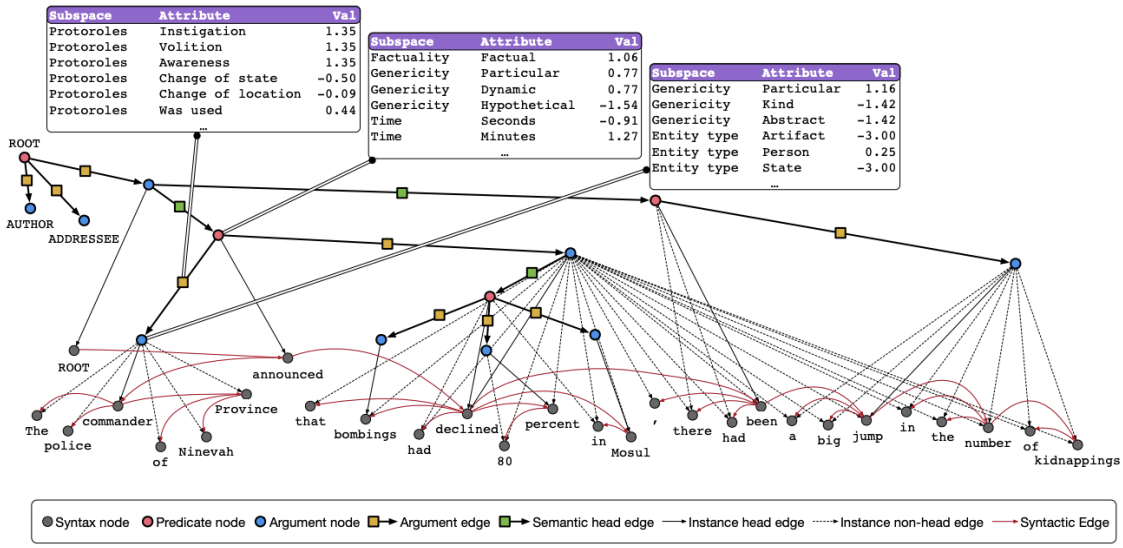

Figure 2: An example Universal Decompositional Semantics graph. Some semantic type information and most syntactic structure information (e.g. dependency relation and part-of-speech tags) are not shown but are available in the dataset (White et al. (2019)).

However, the theory of discrete semantic roles runs into the problem of role fragmentation, which is that linguists cannot agree on the correct number and definitions of these roles that all arguments can be reliably classified into. In other words, there are always edge cases, when using a small number of roles, that could be reasonably classified into multiple categories, and indeed share some combination of all their characteristics. This results in ambiguous boundaries between categories and annotator disagreement. When a large number of roles is used, in order to achieve clear, fine-grained boundaries between categories, many categories are extremely rare and highly similar to more common categories, which raises the question of whether they are indeed distinct categories. In order to fully and unambiguously classify all or almost all feasible arguments, the number of roles required explodes dramatically and this explosion is termed role fragmentation (Reisinger et al. (2015)).

Dowty (1991)'s theory of semantic proto-roles tries to address these concerns by changing the representation of the semantic role of an argument from a discrete category, equivalent to a high-dimensional one-hot vector embedding, to judgments on multiple features. Recent work in Reisinger et al. (2015), White et al. (2016), and White et al. (2019) takes this representations a step further to being a real-valued, lower-dimensional, continuous vector. The individual dimensions of this vector represent an underlying set of semantic properties, or features, of the argument's relationship to the event. These properties are annotated by asking semantic questions about particular arguments in event-bearing sentences. Vector representations for roles allow for a similarity structure to develop, which allows for semantic roles to vary continuously and be similar to multiple discrete semantic roles. Discrete semantic roles emerge from particular configurations of these properties, corresponding to specific regions of the proto-role vector space. When the continuous space of proto-role properties is binned or clustered, the discrete roles emerge, but some are more frequent than others, accounting for the long-tail frequency distribution. Dowty (1991) additionally groups properties into two categories, PROTO-AGENT and PROTO-PATIENT. A PROTO-AGENT property is positive when the argument is more AGENT-like, and vice versa. Dowty's argument is that instead of clearly differentiated discrete semantic roles, there are two prototypical roles, so-called when an argument has more positively-valued properties from one category and less from the other. These two prototypical clusters can be thought of as super clusters, clusters of the clusters that correspond to discrete semantic roles.

The UNIVERSAL DECOMPOSITIONAL SEMANTICS dataset, inspired by the proto-role theory, consists of 1960 sentences with all arguments annotated fully with the 14 proto-role properties in Table 1. Sentences are annotated with dependency parses and semantic parses with predicate

and argument nodes and edges between them, as shown in Figure 2. These nodes and edges are annotated with fine-grained real-valued scalar semantic properties, including proto-role properties, unlike most other semantic parsing datasets. We hypothesize that descriptive statistics will uncover clusters in the proto-role vector space, where exemplars of these clusters correspond to common discrete semantic roles like AGENT, PATIENT, THEME etc. This kind of underlying structure would explain the long tail of rare semantic roles and their similarities to one or more of the common semantic roles. We also hypothesize that clustering this space into 2 super-clusters should uncover PROTO-AGENT and PROTO-PATIENT exemplars, which would be additional verification of Dowty's hypothesis.

The argument for representing semantic roles as vectors with similarity structure has strong parallels to the extremely successful paradigm of using neural networks to learn word embeddings. In both cases, discrete one-hot representations are transformed into continuous vector embeddings where various distance functions, like cosine similarity, Euclidean distance, or dot products, can be used as measures of similarity. The difference is that words are observable random variables, whereas abstract linguistic objects like semantic roles are usually hidden random variables, especially in the task of neural language modeling, and thus cannot be explicitly embedded. Indeed, semantic roles are just one part of a semantic parse tree. While traditionally semantic parse trees are represented as discrete objects, like trees, the neural embedding paradigm raises the question of whether these more complex random variables can also be represented as continuous neural embeddings.

The theory of Tensor Product Representations (TPRs) developed by Smolensky (1990), and the larger framework of Integrated Connectionist/Symbolic Cognitive Architecture (ICS) developed in Smolenksy and Legendre (2006), offers a potential answer to this question. ICS is a comprehensive theory of cognitive computation that is inspired by the recognition that both symbolic computation and neural/connectionist computation contribute necessary but insufficient explanations towards how the mind operates. ICS tries to integrate these two computational systems into one unified architecture that more completely describes cognition. ICS combines symbolic computational properties like discreteness, compositionality, and interpretability with neural computational properties like learning from data, tolerance to imperfections in data, and continuity.

Tensor Product Representations are ICS's solution to the question of how the mind represents data. Symbolic representations are discrete, compositional, and transparent/interpretable structures like trees that are made from a small number of constituent parts. Neural, or connectionist, representations are continuous tensors learned by neural networks that are not clearly compositional or interpretable, but are capable of representing similarities between different objects and embedding and operating on imperfect objects, like sentences with small grammatical errors. TPRs are explicitly-compositional vector embeddings of symbolic structures. At the lowest, formally specified level of analysis, TPRs are simply continuous tensor embeddings. However, TPRs can be described at a higher level as having emergent, discrete structure with consituent parts that can be parsed and queried. TPRs reuse certain tensors in constituent parts for computing the representation of a larger compositional, combinatorial structure. A small number of tensors are combined using the tensor product and sum operations, which corresponds to how symbols can occupy different positions to create constituent nodes and multiple nodes can be combined using edges in parse trees. However, these tensors that are reused can vary slightly to capture soft variations and imperfections in the data. Thus, only a small percentage of the entire continuous tensor space is actually used in computation, resulting in the key TPR property of *approximate discreteness*. The result is that parse trees can be embedded and similarity between them computed.

It is this property of approximate discreteness that corresponds strongly to the proto-role theory. Rather than exact discreteness, the proto-role theory creates continuous vector embeddings, but these embeddings are not uniformly distributed in the vector space and instead create a very narrow and structured space through their clusters.

Recent work has modified the Transformer architecture, and other neural architectures, to be able to compute and learn TPRs of sentences. In this TP-TRANSFORMER, TPRs have a set of role

Proto-role property	Annotator question
instigation	ARG caused the PRED to happen?
volition	ARG chose to be involved in the PRED?
awareness	ARG was/were aware of being involved in the PRED?
sentient	ARG was/were sentient?
change of location	ARG changed location during the PRED?
existed before	ARG existed before the PRED began?
existed during	ARG existed during the PRED?
existed after	ARG existed after the PRED stopped?
change of possession	ARG changed possession during the PRED?
change of state	ARG was/were altered or somehow changed during or by the end of the PRED?
change of state continuous	the change in ARG happened throughout the PRED?
was used	ARG was/were used in carrying out the PRED?
was for benefit	PRED happened for the benefit of ARG?
partitive	only a part or portion of ARG was involved in the PRED?

Table 1: Proto-role property annotations used in this project and the corresponding questions that were asked to annotators (White et al. (2019)).

vectors (role used in a different sense here from semantic roles) that are commonly shared and attended to during computations of embeddings for each token, which are combined with freely generated embeddings, called filler vectors, using either the Hadamard or the Tensor Product. Roles have been shown to learn structural, syntactic information, and fillers content-based, semantic information in previous work. This work has also shown that particular roles can be identified as embedding linguistic categories like predicates, which means that given that a certain role vector is used in embedding a token, then that token will be a predicate with highly probability. The most important result of this research is that, after learning TPRs of highly compositional sequential data using specialized neural architectures, when the role vectors of constituent embeddings of a TPR are experimentally changed and this TPR is transformed back into a sequence, the sequence changes in expected ways with high probability. This is a causal result that shows that the TPR of a compositional symbolic representation of a sequence has the same systematic productive properties as a discrete representation like a parse tree.

Thus, the goal of this project is to investigate whether semantic roles can be interpretably embedded within TPRs in a TP-TRANSFORMER, and whether these emergent representations have the properties of semantic proto-role representations. We hypothesize that, trained on language modeling, without any specific signal guiding the model towards learning semantic role embeddings, the role vectors in the TP-TRANSFORMER assigned to arguments will learn to embed some kind of semantic role structure. For example, one possibility could be that individual learned role vectors correspond to frequent clusters of proto-role properties. Alternatively, the role vectors could embed proto-role properties and attention-weighted combinations of them could represent the overall proto-role vector. We generally wish to investigate whether proto-role representations can be readily interpreted within Transformers and particular properties be associated with particular neural vectors. This work builds on previous interpretability research using neural networks with an inductive bias towards learning TPRs. One major difference is that, since this project is looking to identify annotated semantic data within role vectors, more precise relationships between role vectors and annotations can be quantitatively calculated, instead of the open-ended qualitative interpretation that was more common in previous work.

In Section 2 we provide more detail on the theories of proto-roles and TPRs, and the TPT architecture. Section 3 describes the model training and analysis procedure. Results, visualizations, and analysis of these results are given in Section 4.

2 Background

2.1 Tensor Product Representations

A Tensor Product Representation encodes a constituent in a symbolic structure as a composite of a role, which encodes the structural information (e.g., the dependency relation with another word), and a filler, which encodes the con-

tent of the constituent (e.g., the meaning of a word). Roles are generalizations of positional embeddings, they could be for eg. left-child-of-root, second-position-in-string, or something more complex like a PoS tag. Fillers are the content, they can be thought of as the meanings of words, regardless of their position. The three core operations of a TPR are binding, bundling, and unbinding. Binding is done through the tensor product operation. The binding of a filler to a role creates the representation of a constituent in a symbolic structure. Multiple constituents at the same level of a symbolic structure, such as the same depth in a tree, are bundled together using the sum operation. Fillers can be atoms, vectors that cannot be decomposed further, or TPRs themselves, like a subtree, and can be bound to a role to embed hierarchical compositional structures like trees recursively. Alternatively the information about hierarchical structure can be embedded within roles and fillers can be just embeddings. An important property of the most basic system of TPRs is that role vectors must be orthonormal. The equation for creating a TPR using the sum and tensor product operations is in (1) where S represents the list of symbolic constituents of the structure, which are represented as a pair of the filler and the role.

$$T = \sum_{(i,j) \in S} \hat{f}_i \otimes \hat{r_j} \ (1)$$

$$\tilde{f}_i = T \cdot \hat{r_i} = \sum_{j=1}^{k} (\hat{r_j} \cdot \hat{r_i}) \hat{f}_i \ (2)$$

Importantly, a TPR is interpretable and transparent because simple linear algebraic operations can be used to recover the fillers of any of its constituents. If the dot product of a TPR and particular role vector is taken, then, if the roles are orthonormal, each term in the TPR returns a 0 except for the term with the same role, which returns a 1, and is multiplied with the filler at that position to return the filler. This retrieval is shown in (2). If the roles are not orthonormal, then there is an error term, which can be cleaned up using further operations.

2.2 TP-Transformer

The TP-TRANSFORMER modifies the Transformer architecture to create TPRs of input sentences (Vaswani et al. (2017)). The computation of the fillers is like the computations of traditional contextualized word embeddings in Transformers

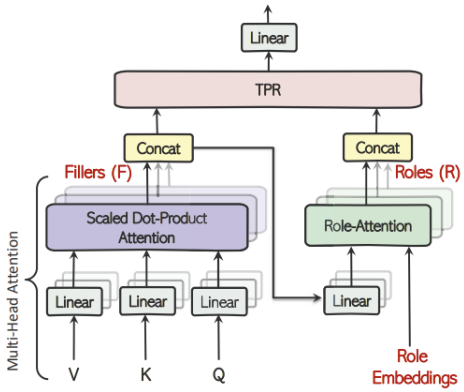

Figure 3: The TP-TRANSFORMER (Jiang et al. (2021)).

at each layer. What this new architecture adds is the learning and computation of role vectors. Each layer contains a dictionary of learned role vectors, creating a role matrix. Each filler vector at each head of each layer is transformed using learned weights and the softmax operation into an attention vector with the same dimensionality as the number of roles in the role matrix. Then, this attention vector is used to compute a blended role vector for each token, which is a weighted average of all roles in the role matrix. Finally, for each token, this blended role vector is bound to the filler vectors using either the Tensor or the Hadamard Product and summed with the unbound filler vector to create a residual connection. The Hadamard Product was shown in previous work to be an optimal lower-rank approximation of the full TPRs (Schlag et al. (2019)).

3 Experimental Procedure

3.1 Language Modeling

We trained a TP-TRANSFORMER with 3 layers, 3 heads per layer, dimension per head 64, model dimension 512, feedforward layer dimension 2048, and 50 roles in the role dictionary of dimension 64 each. The model had a masked language modeling head on top of the encoder, which is linear transformation of each 512-dimensional token embedding into a probability distribution over the vocabulary. In masked language modeling, 15% of tokens are randomly masked in the input, replaced with a dummy token, and must be correctly predicted by the model. Cross-entropy loss between the model's predicted sentence and the actual sentence is calculated as the training signal. The TP-TRANSFORMER is trained from scratch on 600

cluster_id	0	1	2	3	4	5	6	7
instigation	-0.04	1.03	-0.11	0.11	0.09	0.03	-0.12	0.04
change_of_possession	-0.12	-0.36	-0.34	-0.06	-0.18	-0.51	-0.13	-0.17
existed_before	-0.02	1.34	1.34	1.19	-0.16	0.57	1.08	1.38
was_for_benefit	-0.07	0.27	0.04	-0.10	-0.08	-0.17	-0.21	-0.11
change_of_state_continuous	0.03	0.03	0.11	0.04	-0.00	0.09	1.11	-0.00
change_of_state	-0.06	-0.18	-0.08	-0.10	-0.14	-0.15	1.31	-0.12
volition	-0.18	1.27	-0.06	-0.12	-0.12	-0.99	-0.26	-0.15
change_of_location	-0.12	-0.07	0.01	0.12	-0.13	-0.33	-0.02	-0.09
partitive	-0.16	-0.13	-0.25	-0.17	-0.31	-0.25	-0.12	-0.17
existed_during	0.03	1.37	1.37	1.36	1.39	1.30	1.15	1.40
existed_after	0.16	1.28	1.26	1.21	0.37	0.82	0.31	1.15
awareness	-0.16	1.33	0.96	-0.09	-0.01	-1.07	-0.12	-0.09
sentient	-0.39	1.07	1.29	-0.39	-0.28	-1.15	-0.26	-0.27
was_used	0.03	0.73	0.32	1.37	0.39	0.57	0.58	-0.00

Figure 4: Cluster centers with 8 clusters.

MB of text from the C4 dataset for 20 epochs, which consists of text snippets that are of short to medium length. The model is trained using an Adafactor Optimizer (Shazeer and Stern (2018)) with square root learning rate decay and dropout rate of 0.1.

3.2 Clustering of Proto-role Vectors

The K-MEANS clustering algorithm is used to cluster the space of proto-role vectors with random cluster initialization. The optimal number of clusters is determined using the Within-Clusters Sum of Squares method, also known as the elbow method. The clusters are visualized in 2-dimensional space using the t-SNE method. Several frequent words associated with each cluster are qualitatively examined, along with the values of the proto-role properties in the cluster centers, to determine whether the clusters correspond to common discrete semantic roles.

3.3 Probing and Analysis on UDS Dataset

The trained TP-TRANSFORMER is used to encode sentences annotated with proto-role properties in the UNIVERSAL DECOMPOSITIONAL SEMANTICS dataset. These encodings are used to create a dataset of the attention-blended role vector used for argument tokens and the index of the role vector with the highest attention value in the role dictionary of the last layer of the encoder, associated with the corresponding proto-role vector annotation, for each head of the last layer. The blended role vectors are mapped to the proto-role vectors using a linear layer. An accuracy metric is calculated for these probes by comparing the cluster ids of the predicted proto-role vectors with the cluster ids of the actual proto-role vectors. This probe is used to determine which head carries the most proto-role information.

Furthermore, the index of the max-attention role vector is used to predict the cluster ids of its associated proto-role vectors using multinomial logistic regression. The p-values of individual relationships between max-attention role indices and cluster ids are calculated. The same data is also used to predict the continuous real values of each proto-role property using multiple linear regression, with associated p-values.

4 Results and Analysis

4.1 Interpretation of Proto-role Clusters

The elbow method revealed that the best number of clusters was around 8, as shown in Figure 5 in the appendix. It is important to note that there is not a sharp elbow in the plot, which would indicate very cleanly separable clusters. Instead, the soft elbow indicates that the clusters are not clean separations, which is in accordance with the long-tail frequency distribution of fragmented roles, where many rare roles are highly similar to multiple common role categories.

The 8 clusters are highly interpretable and their cluster centers are shown in Figure 4. Cluster 1 could be an AGENT because it is high in all attributes associated with human or living beings,

Cluster Number	Interpretation
Cluster 0	Unclear
Cluster 1	AGENT
Cluster 2	living PATIENT
Cluster 3	INSTRUMENT
Cluster 4	Unclear
Cluster 5	non-living PATIENT
Cluster 6	Object that changes state during event
Cluster 7	LOCATION

Table 2: Qualitative interpretations of cluster centroids of proto-role space.

such as *awareness* and *sentience*, in addition to being high in agentive attributes like *volition* and *instigation*. In contrast, Cluster 2 is high in living attributes, but importantly not in *volition* or *instigation*, which suggests that it could be a living PATIENT. Cluster 3 might correspond to INSTRUMENT, since it is high in *was used*, but negative in living attributes. Cluster 5 could be a PATIENT that is not living because it is strongly negative in living attributes and *volition*. Cluster 6 could be an object that changes state during an event and may cease to exist afterwards, since it is high in *change of state* and *change of state continuous*. Cluster 7 is difficult to interpret solely based on the cluster centers, but common words assigned to that cluster reveal that it is highly associated with locations. Clusters 4 and 0 are difficult to interpret. Cluster 0 seems to be a default cluster since it is close to 0 for all attributes.

Clustering of this space into 2 super-clusters very clearly reveals a PROTO-AGENT cluster and a PROTO-PATIENT cluster. One of the clusters is very positive in highly agentive attributes such as sentience, awareness, volition, and instigation, whereas the other is moderately negative in these attributes. These results are shown in Figure 8 in Appendix A.

In conclusion, these results largely follow our hypothesis about the proto-role space, that this space is weakly clustered and common clusters correspond to common discrete semantic roles.

4.2 Associations between TPT Roles and Proto-roles

The results of mapping attention-blended role vectors of argument tokens to their annotated proto-role vectors shows that the roles of Head 3 were most likely to contain information about proto-roles (see Table 3). Thus, we show the results from that head.

Input source	Accuracy
Head 1	14.54%
Head 2	17.73%
Head 3	24.46%
Concatenated Heads	22.00%
All Heads	16.10%

Table 3: Accuracy of linear classifier mapping attention-blended role vectors to proto-role vectors. All heads indicates that the vector from each head paired with the same proto-role vector was a separate data point used to train the same classifier.

In using linear and logistic regression to predict clusters and individual proto-role properties from the index of the role with the highest attention value, we find that results are significant for most associations between clusters and role indices, but not properties and role indices (although there are still a few significant coefficients in that case that are useful as supplemental evidence for some interpretations). This may just be because linear regression is a more difficult task than logistic regression. Instead of predicting the exact scalar values, logistic regression is predicting a general blob of proto-role vectors, within which there is high variation in the exact values. This does however show that there is not a clear relationship between individual role vectors and proto-role properties, but there are some relationships where certain role vectors can be interpreted as embedding certain semantic roles. In the following paragraphs, we summarize the most interpretable results. Figure 6 shows the significant coefficients for predicting clusters and Figure 11 shows the significant coefficients for predicting individual properties. For reference, Table 2 shows the qualitative interpretations for each cluster.

Role 3 positively predicts clusters 5 and 6 and

cluster_id	R3	R6	R8	R10	R15	R19	R20	R21	R25	R29	R31	R32	R34	R35	R36	R38	R40	R42	R43	R44	R47
1	1.21	-33.21	0.14	25.55	-31.02	3.19	3.12	4.66	0.32	-1.40	2.20	-27.33	-0.83	17.57	1.61	18.58	-9.65	-0.18	-0.71	2.06	1.70
2	-17.43	-26.77	4.42	-3.35	-22.72	6.84	7.49	8.95	5.88	-7.09	6.51	-20.01	5.04	21.80	-16.82	24.34	-5.80	-19.84	5.03	6.41	5.93
3	-21.92	0.62	0.79	-4.60	-27.96	2.41	0.80	2.93	1.48	-0.10	0.99	-25.49	1.42	-4.89	1.31	20.07	23.78	1.72	1.09	1.16	1.37
4	-22.49	-32.35	4.12	-4.66	-27.65	5.29	3.65	5.85	4.57	3.72	3.74	-26.04	3.88	17.56	3.90	21.97	-8.01	3.21	3.95	3.75	4.22
5	7.08	-21.43	5.61	-1.17	-13.94	-9.37	5.88	-3.91	6.20	5.57	6.28	-14.20	6.50	-2.19	7.08	-5.47	-2.72	6.39	6.53	4.04	6.15
6	6.49	-24.32	5.03	-1.39	-15.98	5.81	4.42	-5.55	4.77	3.89	4.26	-16.56	5.58	-2.90	-16.71	-6.72	28.97	5.80	5.24	5.40	5.51
7	-16.93	-25.82	6.35	-2.57	-20.06	7.40	6.56	8.38	7.30	5.85	6.97	7.69	7.04	-2.47	-16.63	-7.80	-4.98	-20.38	6.85	6.78	7.16

Figure 5: Coefficients of multinomial logistic regression predicting cluster of proto-role vector from the index of the max-attention role used. Cluster 0 is not shown because it was used as the reference.

cluster_id	R3	R6	R8	R10	R15	R19	R20	R21	R25	R29	R31	R32	R34	R35	R36	R38	R40	R42	R43	R44	R47
1	False	True	False	True	True	True	True	True	False	True	True	True	True	True	False	True	False	False	False	True	True
2	True	False	False	True	True	False	False	False	False	False	False	True	False	True	True	True	True	True	False	False	False
3	True	False	True	False	True	True	True	True	True	False	True	True	True	True	False	True	True	False	True	True	True
4	True	False	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True
5	True	False	True	False	True	True	True	True	True	True	True	True	True	True	True	True	False	True	True	True	True
6	True	True	True	False	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True
7	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True	True

Figure 6: P-values of the multinomial logistic regression coefficients.

negatively clusters 2, 3, 4, and 7 (see Figure 5). It also positively predicts the *change of state* properties and negatively the *existed before* property (see Figure 10). This implies that Role 3 encodes properties that are common between clusters 5 and 6, which are both non-living PATIENT-like clusters.

The most significant result is that role 10 is very clearly representing an AGENT semantic role. It is strongly predictive of cluster 1 and slightly negatively of all others (see Figure 5).

Role 15 is highly negatively predictive of cluster 1, 3, and 4 (see Figure 5). It is also negatively predictive of *volition*, *existed before*, and *existed during* (see Figure 10). It can be clearly said that it is encoding the opposite of AGENT-like traits, but it is difficult to ascertain what it is positively encoding.

Role 19 is negatively predictive of cluster 5, and moderately positively predictive of all others (see Figure 5). This suggests that it encodes a property that all clusters share except for cluster 5. Cluster 5 is most negative in living attributes and volition, which suggests that this role encodes the opposite of those properties.

Role 32 is another clearly interpretable embedding. It seems to encode locations, since it is positively predictive of cluster 7 and strongly negatively of all others (see Figure 5). It is also negatively predictive of *volition* and *awareness* (see Figure 10), which further strengthens this hypothesis.

Role 35 is strongly predictive of clusters 1, 2, and 4, and negatively predictive of clusters 3, 5, 6, and 7 (see Figure 5). The properties shared by clusters 1 and 2 are that they are high in living attributes. On the other hand, clusters 3, 5, 6, and 7 are low in living attributes. Thus, Role 35 can be interpreted as a configuration of the living attributes.

Role 38 is strongly predictive of clusters 1, 2, 3, and 4 and negatively of clusters 5, 6, and 7 (see Figure 5). Clusters 1 and 2 are high in living attributes, whereas clusters 5, 6, and 7 are objects that are high in PROTO-PATIENT properties.

One reason why Roles 10 and 32 stand out as being clearly associated with one cluster over all others is that AGENT and LOCATION are highly distinctive semantic roles that only very specific types of objects can play. On the other hand, many more types of objects can be involved in events in a PROTO-PATIENT capacity.

Finally, Roles 6, 8, 20, 21, 25, 29, 31, 34, 36, 40, 42, 43, 44, and 47 are difficult to qualitatively interpret because they either have too many insignificant coefficients, or they are too weakly predictive of clusters, or it is difficult to find a meaningful commonality from which clusters they are meaningfully predictive of.

5 Conclusion and Future Work

In this work, we posit a conceptual link between the theory of proto-roles and the Integrated Connectionist/Symbolic Cognitive Architecture, particularly the motivation behind Tensor Product Representations. Following intepretability research that uses neural network architectures modified to learn TPRs of individual words in sentences, we train a TP-TRANSFORMER on part of the C4 dataset and use this language model to encode UNIVERSAL DECOMPOSITIONAL SEMANTICS sentences annotated with proto-role properties. We cluster the space of proto-role vectors to uncover clusters that correspond to common discrete semantic roles, further validating the theory. Finally, we analyze significant associations between role vectors in the TP-TRANSFORMER and proto-role properties and clusters, uncovering some interpretable encodings within role vectors.

References

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Yichen Jiang, Asli Celikyilmaz, Paul Smolensky, Paul Soulos, Sudha Rao, Hamid Palangi, Roland Fernandez, Caitlin Smith, Mohit Bansal, and Jianfeng Gao. 2021. Enriching transformers with structured tensor-product representations for abstractive summarization. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Vasin Punyakanok, Dan Roth, and Wen tau Yih. 2008. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, 34(2).

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. 2019. Enhancing the transformer with explicit relational encoding for math problem solving. *CoRR*, abs/1910.06611.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.

Paul Smolenksy and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT Press, Cambridge, MA, US.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, et al. 2019. The universal decompositional semantics dataset and decomp toolkit. *arXiv preprint arXiv:1909.13851*.

A Additional Figures

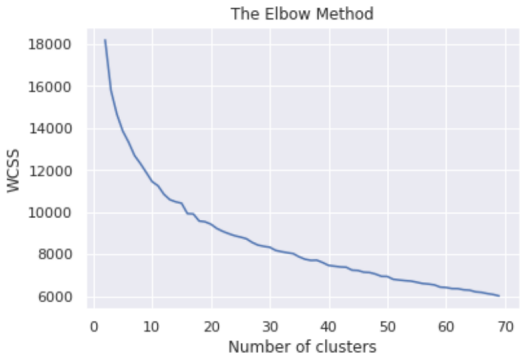

Figure 7: Plot of WCSS as a function of number of clusters used in K-means with random initialization.

Cluster	0	1
instigation	0.03	0.55
change_of_possession	-0.17	-0.35
existed_before	0.61	1.33
was_for_benefit	-0.12	0.19
change_of_state_continuous	0.11	0.08
change_of_state	0.00	-0.12
volition	-0.25	0.71
change_of_location	-0.08	-0.05
partitive	-0.21	-0.17
existed_during	1.09	1.37
existed_after	0.69	1.26
awareness	-0.20	1.17
sentient	-0.41	1.13
was_used	0.46	0.57

Figure 8: Cluster centers with 2 clusters, showing proto-Agent and proto-Patient clusters.

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
it	I	I	letter	nothing	meal	plans	India
thing	They	me	report	problem	suggestion	eggs	place
what	you	you	card	work	something	prices	country
that	We	We	draft	service	them	waters	name
I	He	them	dog	reason	dust	resources	you
nothing	she	Bush	opinion	food	concern	missile	food
They	me	people	items	place	car	process	town

Figure 9: Exemplar words from each cluster.

proto-role property	R3	R6	R8	R10	R15	R19	R20	R21	R25	R29	R31	R32	R34	R35	R36	R38	R40	R42	R43	R44	R47
instigation	0.05	-0.53	-0.12	-0.32	-0.46	0.12	0.16	0.31	-0.32	-0.36	0.07	-0.47	-0.43	0.24	0.14	-0.46	-0.55	-0.28	-0.36	0.06	-0.04
change_of_possession	-0.18	0.18	0.02	0.23	0.13	-0.03	-0.00	-0.03	0.04	0.10	-0.06	0.15	-0.00	-0.05	0.14	-0.07	0.79	0.03	0.08	-0.01	-0.03
existed_before	-0.72	-1.05	-0.59	0.23	-1.46	-0.06	0.08	0.08	-0.47	-0.86	-0.09	-0.24	-0.39	0.18	-0.52	-0.28	0.15	-0.24	-0.52	-0.07	-0.22
was_for_benefit	0.03	0.00	0.05	0.21	0.07	0.14	0.27	0.24	0.09	-0.08	0.14	0.11	0.01	0.29	0.43	-0.39	-0.02	-0.33	0.05	0.10	0.09
change_of_state_continuous	0.56	-0.02	0.04	0.04	-0.03	0.05	0.03	-0.01	0.05	0.08	0.03	0.00	0.12	0.08	-0.03	0.21	0.64	0.15	0.10	0.07	0.07
change_of_state	0.55	0.21	0.23	0.33	0.13	0.15	0.18	0.17	0.16	0.18	0.14	-0.21	0.31	0.19	0.23	0.05	0.40	0.39	0.23	0.18	0.21
volition	-0.39	-0.74	-0.62	0.75	-2.23	-0.06	0.10	0.15	-0.69	-0.78	-0.16	-1.18	-0.83	0.28	-0.22	-1.17	-1.89	-0.61	-0.83	-0.16	-0.36
change_of_location	0.20	0.21	0.14	0.37	-0.05	0.10	0.18	0.14	0.15	0.20	0.14	0.21	0.14	0.22	0.35	0.25	0.21	0.20	0.18	0.22	0.18
partitive	-0.06	0.01	-0.07	-0.14	-0.21	-0.07	-0.10	-0.02	-0.14	-0.03	-0.06	-0.29	-0.12	-0.12	0.04	-0.13	-0.70	-0.25	-0.14	-0.14	-0.07
existed_during	-0.33	-0.84	-0.30	-0.01	-1.37	-0.09	-0.08	-0.04	-0.21	-0.51	-0.16	-0.48	-0.28	-0.02	-0.18	-0.07	-0.09	-0.30	-0.29	-0.15	-0.17
existed_after	-0.17	-0.68	-0.56	0.16	0.08	-0.22	-0.05	-0.09	-0.37	-0.88	-0.17	-0.74	-0.44	-0.01	-0.43	-0.13	-1.28	-0.59	-0.49	-0.18	-0.29
awareness	-0.39	-0.99	-0.82	0.49	-1.13	-0.21	0.14	0.06	-0.69	-1.03	-0.15	-1.01	-0.94	0.30	-0.48	-0.50	-0.79	-0.92	-0.93	-0.15	-0.40
sentient	-0.31	-0.67	-0.69	-0.58	-0.79	-0.02	0.49	0.44	-0.48	-0.90	0.11	-1.05	-0.81	0.69	-0.51	0.07	-0.46	-0.71	-0.76	0.09	-0.25
was_used	-0.44	-0.27	-0.22	0.68	0.61	-0.11	-0.06	0.04	-0.27	-0.37	-0.12	-0.33	-0.27	-0.16	0.13	0.18	0.60	-0.11	-0.29	-0.19	-0.17

Figure 10: Coefficients of multiple linear regression predicting proto-role properties from the index of the max-attention role used.

proto-role property	R3	R6	R8	R10	R15	R19	R20	R21	R25	R29	R31	R32	R34	R35	R36	R38	R40	R42	R43	R44	R47
instigation	False	False	False	False	False	False	False	False	False	True	False	False	True	False	False	False	False	False	True	False	False
change_of_possession	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
existed_before	True	True	True	False	True	False	False	False	True	True	False	False	True	False	False	False	False	False	True	False	False
was_for_benefit	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False	False	False	False	False
change_of_state_continuous	True	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
change_of_state	True	False	True	False	False	False	False	False	False	False	False	False	True	False	False	False	False	True	True	False	False
volition	False	False	True	False	True	False	False	False	True	True	False	True	True	False	False	True	True	True	True	False	True
change_of_location	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False
partitive	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	False	False
existed_during	False	True	True	False	True	False	False	False	False	True	False	False	True	False	False	False	False	False	True	False	False
existed_after	False	False	True	False	False	False	False	False	True	True	False	False	True	False	False	False	True	True	True	False	False
awareness	False	True	True	False	False	False	False	False	True	True	False	True	True	False	False	False	False	True	True	False	True
sentient	False	False	True	False	False	False	True	False	True	True	False	False	True	True	False	False	False	True	True	False	False
was_used	False	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False	False

Figure 11: P-values of the multiple linear regression coefficients.