Emergent Semantic Proto-role Structure in Tensor Product Representations

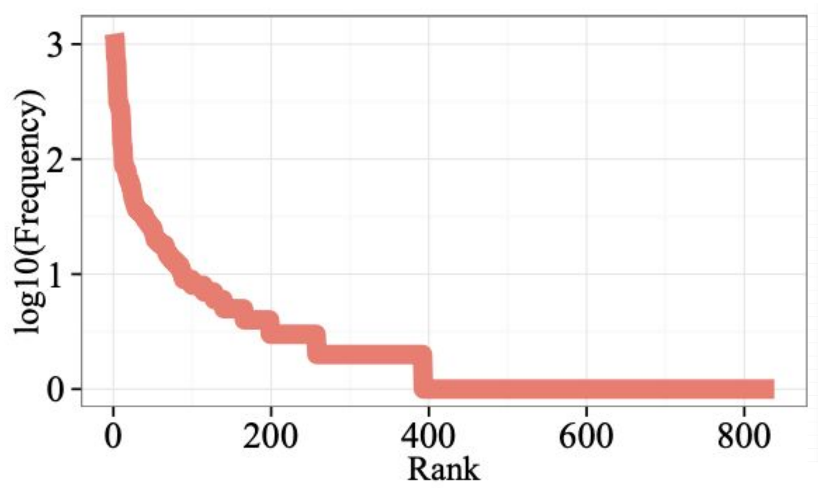
Devanshu Singh, Paul Smolensky

Data Science Capstone Project: Spring 2022

Johns Hopkins University

Introduction and Hypothesis

- The goal is to investigate whether semantic proto-role representations emerge within neural representations in Transformers and can be readily interpreted.
- Based on Dowty's linguistic representation theory of semantic proto-roles as being a more accurate form of representation for semantic roles than discrete semantic roles.
- Semantic roles are the roles that an argument, usually noun, plays in an event. usually verb. Events usually have multiple semantic roles.
- Discrete semantic roles not able to fully explain semantic role information and run into the problem of role fragmentation. Role fragmentation is that experts can't agree on the correct number of different role categories.
- Frequency distribution of discrete semantic roles follows a long tail distribution.



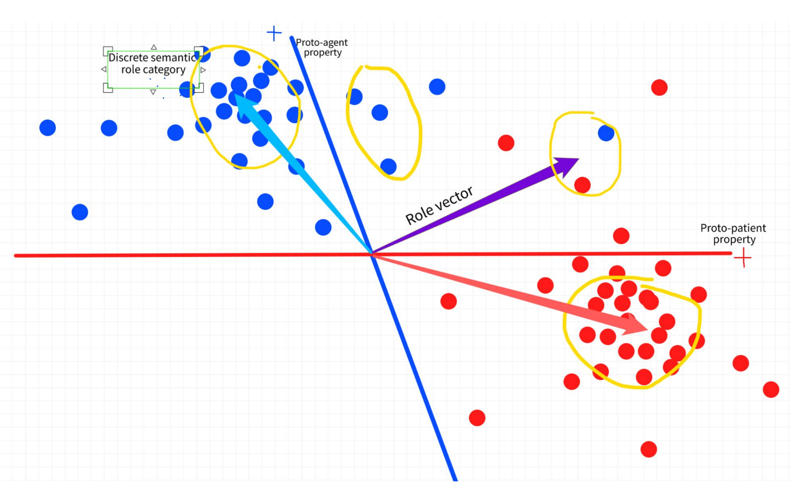


Figure 1: Long tail Log frequency distribution of discrete semantic role clusters of proto-role space (Left). Visual aid to visualize what the space of proto-role vectors looks like and how the clusters correspond to discrete semantic roles.

- Semantic proto-roles create continuous vector representations for semantic roles. Discrete semantic roles emerge from particular clusters of these properties. The tails of these clusters account for rarer semantic roles.
- The Universal Decompositional Dataset contains 1960 sentences which are annotated with a syntactic and semantic graph, and particular arguments are annotated with the full vector of proto-role properties.
- <u>We hypothesize that descriptive statistics will uncover prototypical clusters in this space, where exemplars of these clusters correspond to common discrete semantic roles like Agent, Patient, Theme etc.</u> 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.

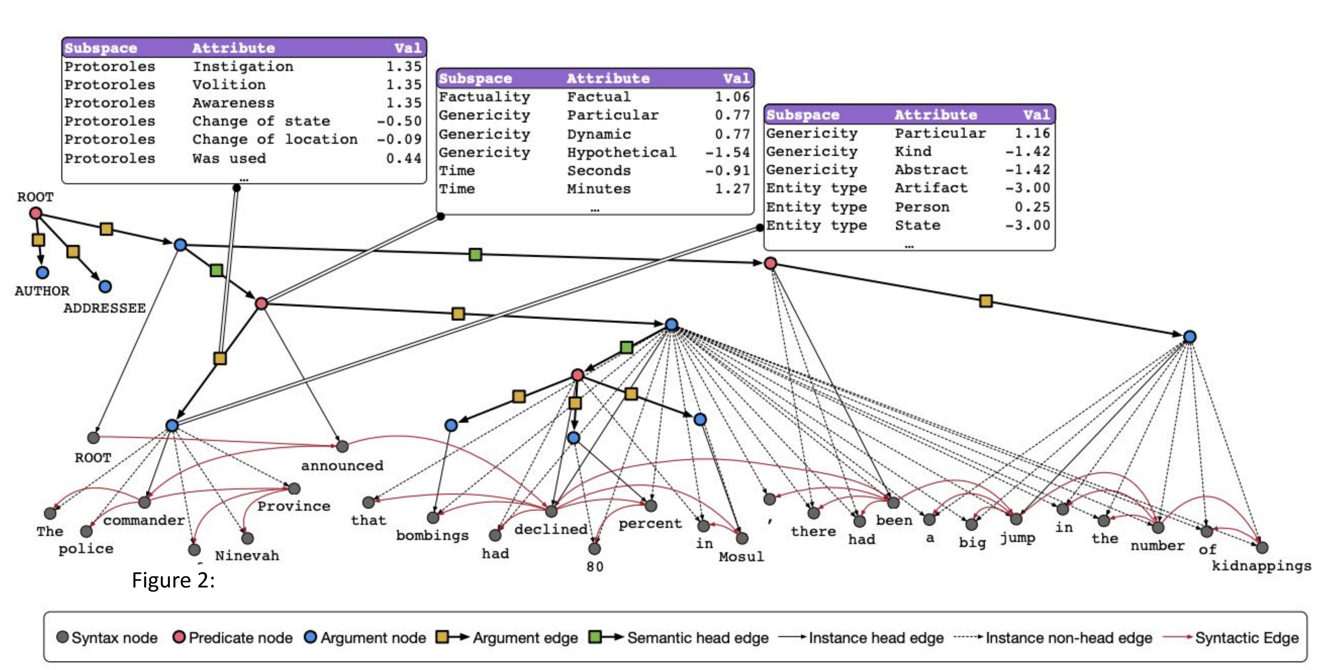


Figure 2:

Figure 1: 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.

References

1. Yichen Jiang, Asli Celikyilmaz, Paul Smolensky, Paul Soulos, Sudha Rao, Hamid Palangi, Roland Fernandez, Caitlin Smith, Mohit Bansal, Jianfeng Gao. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* June 2021. Published by Association for Computational Linguistics.
2. Reisinger, Dee Ann, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic Proto-roles. *Transactions of the Association of Computational Linguistics* 3: 475-488.
3. White, Aaron Steven, Dee Ann 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*, 1713–1723. Austin, Texas: Association for Computational Linguistics.

Tensor Product Representations and TPTs

- Dowty's theory corresponds to Smolensky's theory of Tensor Product Representations, which are explicitly-compositional vector embeddings of symbolic structures. TPRs aim to represent discrete structures like trees as continuous neural vectors, but still be able to uncover interpretable structures.
- These TPRs are sums of tensor product bindings between role vectors, which represent structural, syntactic information, and symbol vectors, which represent semantic information.

king - queen = (royal : status + male : gender) - (royal : status + female : gender)

man - woman = (commoner : status + male : gender) - (commoner : status + female : gender)

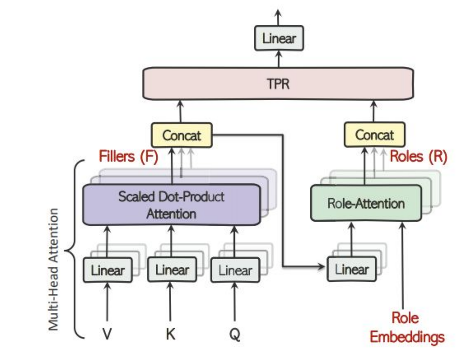


Figure 3: Tensor product of common word vector analogies with structural roles and filler symbols (Left). The Tensor Product Transformer architecture (Right).

- The Tensor Product Transformer (TPT) tries to learn TPRs of sentences
- <u>We hypothesize that, trained on language modeling, the role vectors in the TP-Transformer assigned to arguments will learn to embed some kind of semantic role structure.</u> We hope that individual learned role vectors correspond to frequent clusters of proto-role properties. We generally wish to investigate whether proto-role representations can be readily interpreted within Transformers and particular properties be associated with particular neural vectors.

Procedure

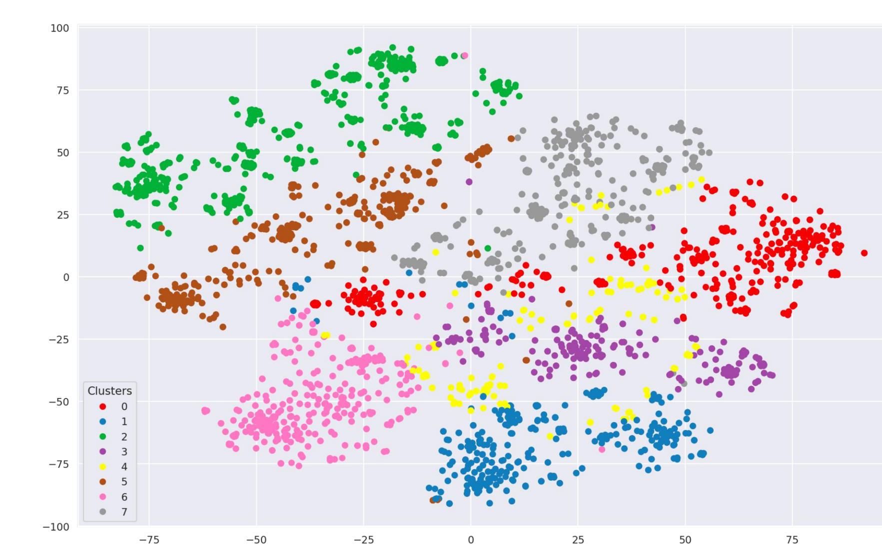
- Trained TPT language model (TPT-LM) on 600 MB of text data for 20 epochs.
- Use TPT-LM to encode UDS dataset with annotated protoroles. Create dataset of role vectors learned by model for each token associated with that token's protorole vector annotation.
- Cluster the space of proto-role vectors using k-means clustering. Examine clusters to see if they are prototypical of certain common discrete semantic roles.
- Probe for proto-role information among role vectors by using regression to map TPT role vectors to proto-role vectors and their clusters. Report p-values.
- Analyze and test for associations between particular roles used for particular tokens and clustered proto-roles and individual proto-role properties.

Discussion

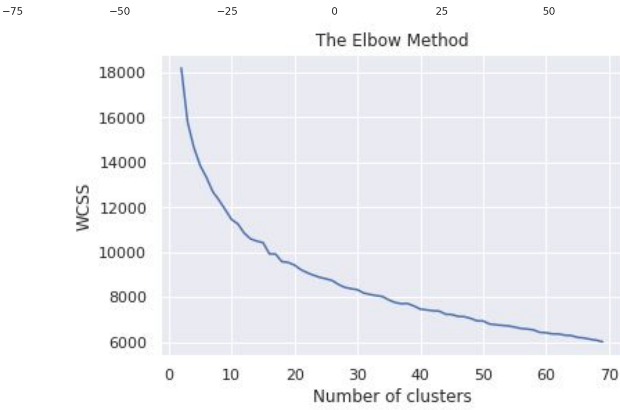
1) Hypothesis 1 seems to be correct and confirms previous results that the protorole space is weakly clustered.
   a) Cluster 1 might correspond to Agent, Cluster 3 might correspond to Instrument, Cluster 5 might correspond to a Patient that is not living, Cluster 2 might correspond to a living Patient, Cluster 6 seems to be an object that might change state often in its role in events. Cluster 7 encode locations.
2) TPT roles are much more predictive of protorole clusters than protorole values themselves. This may be because linear regression is a more difficult prediction task or it may be that role vectors encode higher level information, above the subsymbolic level of protorole properties. However, there are only some clearly identifiable role vectors acting as common semantic roles, especially Role 10.
   a) Role 3 positively predicts clusters 5 and 6, and the change of state properties and negatively the existed before property. Role 10 strongly predicts cluster 1, and nothing else, meaning that it might be cleanly identifiable as the Agent role. Role 40 is strongly predictive of a role that does not have volition, but changes state and possession, which seem to be attributes that both cluster 3 and 6 share.

Results

- Focus on the role vector from dictionary with the highest attention value for analysis. Showing results for the head that seems to be most predictive of protorole vectors.
- Probing accuracy with all head roles: 16.1%, Probing accuracy with concatenated roles: 22.00%, Probing accuracy with head 3 roles: 24.46%, Probing accuracy with head 1 roles: 14.54%, Probing accuracy with head 2 roles: 17.73%.
- Using Within-Cluster Sum of Squares method, 8 seems to be the optimal number of clusters in the protorole space.
- For regression, cluster 0 was used as a reference since it seems to be a default cluster not predictive strongly of proto-role properties.



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



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 4: T-SNE plot of clustered proto-role vectors (Left). Cluster centers and their corresponding property values (Right). Plot of WCSS as a function of number of clusters used in K-means with random initialization (Below Left). Some common words in each cluster (Below Right)

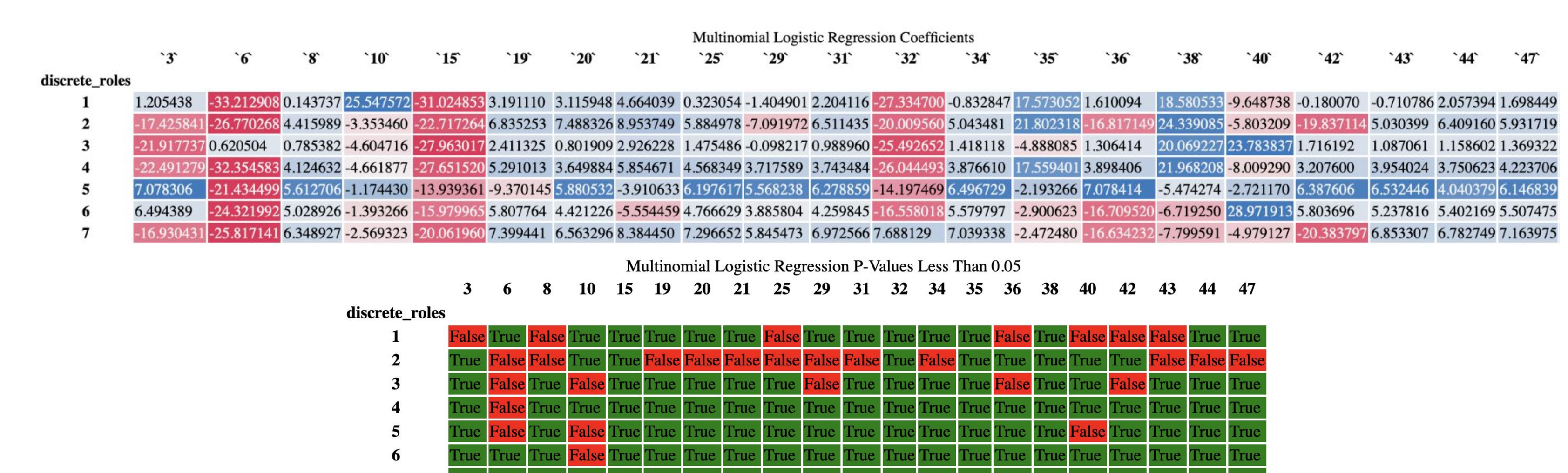


Figure 5: Multinomial logistic regression coefficients with clustered role as a function of role vector with max attention (Above). P-values below 0.05 for these coefficients (Below).

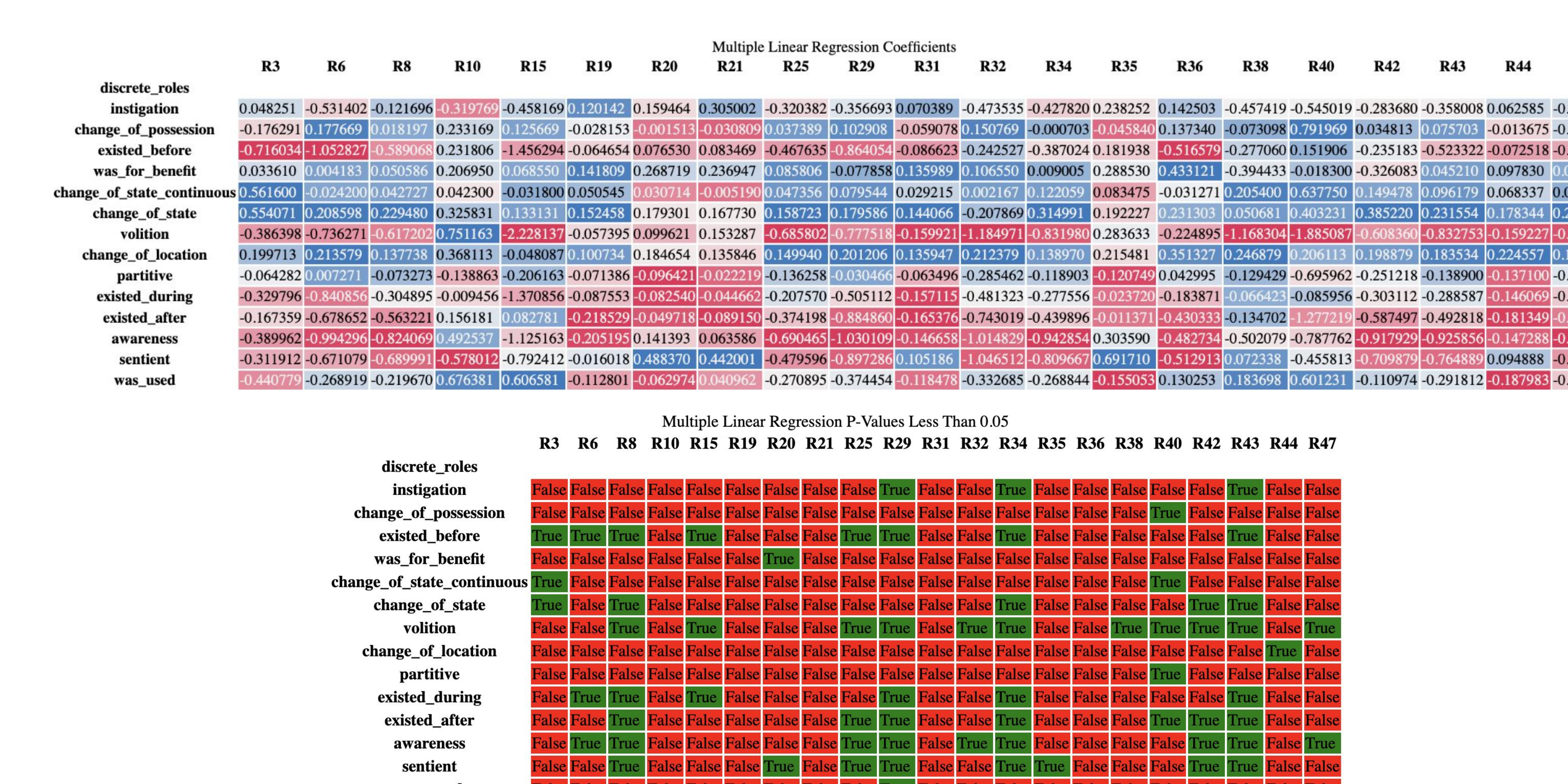


Figure 6: Multiple linear regression coefficients with protorole property value as a function of role vector with max attention (Above). P-values below 0.05 for these coefficients (Below).