Semantic Role Circuits in Language Models

Team: Automated Null Space Projection (ANLP)

Druhan Rajiv Shah
IIIT Hyderabad

Sidharth K
IIIT Hyderabad

Anshul Krishnadas Bhagwat
IIIT Hyderabad

Abstract

The field of Mechanistic Interpretability (MI) aims to determine how information from the input is used by the internals of transformer-based language models. Several breakthroughs have been made in this field, but most of these are either on purely English-based models, or on multilingual models and their common linguistic abilities. This project aims to explore semantic role circuits and their functioning in monolingual models and contrast their structure across models traind on morphologically varied languages.

1 Introduction

The advent of large-scale generative language models has revolutionized natural language processing (NLP), yet their internal workings remain largely opaque. This "black box" problem is a significant barrier to building truly trustworthy and reliable AI systems. The field of Mechanistic Interpretability (MI) seeks to address this by reverse-engineering the specific, human-understandable algorithms that models learn during training (Elhage et al., 2021).

However, a lot of MI research has been concentrated on English-language models, creating a critical gap in our understanding of how a Transformer-based model, adapts to the typological diversity of human languages. Different languages encode grammatical relationships in fundamentally different ways. For instance, English (Germanic) relies heavily on a fixed Subject-Verb-Object (SVO) word order . In contrast, Hindi (Indo-Aryan) uses a more flexible but primarily Subject-Object-Verb (SOV) order (Verma, 1970), marking grammatical roles with distinct postpositional particles. Telugu (Dravidian) is also SOV but employs a highly agglutinative, morphological case system, where roles are marked by suffixes attached directly to nouns.

This research proposes a controlled, cross-linguistic experiment to investigate how these typological differences shape the internal circuits of language models. We will train three identical, autoregressive models (~124.5M parameters) from scratch, one each on English, Hindi, and Telugu. We will then use visualisations, probes and causal interventions to identify and compare the neural circuits each model develops to process semantic roles.

By comparing the circuits that emerge in response to these distinct grammatical strategies, we can move beyond simply knowing that models work for different languages to understanding precisely how they adapt their computational mechanisms. This study will attempt to provide a mechanistic, comparative analysis of semantic role processing in generative models across different language families, offering foundational insights into the functioning of Transformer-based AI systems.

2 Background and Prior Work

Semantic Role Labeling (SRL) SRL is a central process in NLP that seeks to uncover the roles that words (or their computational counterparts: tokens) perform as arguments to the sentence's core verb or predicate. The PropBank (Palmer et al., 2005) schema provides a set of roles that each argument is classified into, including `ARG0` (the entity performing the action out of their own volition), `ARG1` (the entity on whom the action is performed), and `ARGM-LOC` (the verb modifier that indicates location) among others. While classifier models trained to classify tokens by semantic role are numerous, we aim to study generative models and their use of semantic roles instead.

The representation of semantic roles in each language is unique, with English relying on word order, Hindi using distinct post-positional case markers, and Telugu (which is agglutinative) using case-specific suffixes to indicate, but not necessarily determine semantic roles (Vaidya et al., 2011).

Mechanistic Interpretability (MI) MI is a field of study that attempts to reverse-engineer the internal workings of neural models, particularly by obtaining insights about the pathways that information takes in the process of generating outputs. These pathways are isolated in the form of transformer *circuits*: a narrow, functional subgraph of model activations and parameters that perform a single specific task (Elhage et al., 2021). In our case, we seek a circuit that uses the semantic roles of the previous tokens in order to narrow down the sample space for the next generated token.

Identifying these circuits runs the risk of drawing spurious conclusions from observed correlations, however. In order to determine the causality of these observed circuits, activations are *intervened* upon and ablated to observe their effect on the output. This practice has been instrumental in identifying key mechanisms like Induction Heads, and has yielded useful linguistic insights such as the circuits used in Indirect Object Identification.

3 Methodology and Timeline

This project aims to cover two phases: the training and evaluation phase, and the interpretability phase. The first, while theoretically simple, will require some time since we seek to train three GPT-like (Radford et al., 2019) models from scratch on the task of Next Token Prediction. We expect to be done with this by the Mid-submission.

In the second phase, we intend to probe the models' layers and attention heads for semantic role information, test their validity through causal interventions, and finally reconstruct circuits that use prior semantic roles to generate tokens. Intermediate results can include verification of Induction Heads and head ablations.

Comparison of circuits between languages will be qualitatively done, since linguistic nuance is to be expected when dealing with such morphologically distinct languages as the chosen three.

3.1 Datasets

The datasets we will use for training models on each language are their respective Wikipedia datasets (Foundation, 2023). These are sufficiently large and varied to train autoregressive generation on. For SRL datasets, we use the respective PropBanks (Palmer et al., 2005; Bhat et al., 2017; Jindal et al., 2022; Akbik et al., 2015).

References

Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. Generating high quality proposition Banks for multilingual semantic role labeling. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia. 2017. *The Hindi/Urdu Treebank Project*, pages 659–697. Springer Netherlands, Dordrecht.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Wikimedia Foundation. 2023. Wikimedia downloads.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ashwini Vaidya, Jinho Choi, Martha Palmer, and Bhuvana Narasimhan. 2011. Analysis of the Hindi Proposition Bank using dependency structure. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 21–29, Portland, Oregon, USA. Association for Computational Linguistics.

Shivendra Kishore Verma. 1970. Word order in hindi. *Archiv orientalni*, 38:28–32.