Project Mid Submission

Team: Automated Null Space Projection (ANLP)

Druhan Rajiv Shah
IIIT Hyderabad

Sidharth K
IIIT Hyderabad

Anshul Krishnadas Bhagwat
IIIT Hyderabad

Abstract

This report details the progress on our investigation into cross-linguistic semantic role circuits. As outlined in our proposal, the first phase of this project involved training three monolingual, autoregressive transformer models for English, Hindi, and Telugu. We have started this phase and have begun conducting some initial exploratory experiments. We have also conducted elementary tests on a pretrained `gpt2-small` model for English, and are in the process of training the Hindi and Telugu models. Preliminary analyses have been conducted to check for the presence of mechanisms to account for morphosyntactic features that each language requires.

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 ($\sim$ 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.

Background and Related Work

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 postpositional case markers, and Telugu (which is agglutinative) using case-specific suffixes to indicate, but not necessarily determine semantic roles (Vaidya et al., 2011). The work by Ghosh et al. relies on the well-established idea that languages which have no Dominant word order (like Hindi and to an extent Telugu) establish information about semantic roles in sentences with morphosyntactic features like *kāraka* markers or *vibhaktī* affixes (Vaidya et al., 2011). Thus, models can be expected to use differing methods to encode and use information like semantic roles in order to generate tokens.

The field of Mechanistic Interpretability (MI) seeks to reverse-engineer the internal algorithms learned by transformer-based language models during training. This approach moves beyond performance metrics to explain the specific computational mechanisms underlying a model's behavior. The fundamental unit of analysis in MI is the transformer circuit: (Elhage et al., 2021) a subgraph of model parameters and activations responsible for a discrete task. Research in this area has successfully identified key circuits, such as Induction Heads (Olsson et al.), and circuits for indirect object identification (Wang et al.), which are critical for in-context learning and recall. However, a significant portion of MI research has concentrated on English-language models. This focus creates a critical gap in understanding how a transformer's architecture adapts to the vast typological diversity of human languages. Grammatical relationships are encoded in fundamentally different ways across language families, and the circuits developed to process them are likely to differ accordingly.

Experimental progress

Model Training

As per our proposal, we chose to train three identical, autoregressive models (~124.5M parameters) from scratch, one each on English, Hindi, and Telugu. The Hindi and Telugu models are currently in the process of being trained with the English model to immediately follow. The models will use the GPT2 architecture (Radford et al., 2019).

Datasets

We use the Wikipedia dump (Foundation, 2023) as our training corpus. The raw data was pre-processed and tokenized using a Byte-Pair Encoding (BPE) tokenizer trained on the corpus, with a vocabulary size of 50,257. The datasets for semantic-role information for future experiments is going to be the PropBanks for the respective languages (Palmer et al., 2005; Bhat et al., 2017; Jindal et al., 2022; Akbik et al., 2015).

Preliminary Analysis

While awaiting the completion of training for the Hindi and Telugu models, we have begun our investigation into a baseline English model, which in this case is `openaicommunity/gpt2` which we shall call `gpt2-small`. Our initial analysis focuses on identifying attention heads that are sensitive to syntactic and semantic dependencies, which are precursors to full semantic role circuits.

A exploratory test showed that `gpt2-small` in English depends very heavily on Positional Encodings, and similar positional information. This is as expected since English encodes semantic role information through word order. Indeed a similarly exploratory test may be done to analyse the same model trained without the use of positional embeddings at all in order to analyse the work of Haviv et al. where such models perform near-identically.

Standard visualization techniques [1] were employed to analyze attention patterns on sample sentences. We observed the emergence of several specialized heads, including:

- *Previous Token Heads*: These heads attend strongly to the immediately preceding token, a fundamental mechanism in autoregressive models.

- *Induction Heads*: As documented in prior work, we identified heads that appear to implement a very barebones form of in-context learning, completing patterns like

[1] See `Code/shenanigans/analysis.ipynb` in the repository.

`AB...A -> B`. These are critical for copying and recall mechanisms.

- *Semantic Role Depepndency Heads:* Notably, we see that `gpt2-small` in English does have heads that attend to semantic role-based dependencies (*e.g.* Pred → ARG0) but these heads are either not causal, or they do not hold up to sentences which have those same semantic roles placed further apart. The Verb→Subject circuit is the strongest found, with 86 specialist heads. This circuit is highly distributed; ablating the top 3 heads combined could change only about 26-46% of the prompts. However, a complete ablation of all 86 heads was catastrophic, resulting in totally different outputs (average KL Divergence of 0.5677), proving the collective circuit is causally essential [2].

Next Steps

Our immediate priority is to complete the training of the language-specific models. Once training is complete, we will proceed with the second phase of our project as outlined in the proposal:

1. Training linear probes on the internal representations of all three models to determine where and how explicitly semantic role information is encoded.

2. Using activation patching to causally trace the flow of information related to semantic roles (e.g., from a noun marked as an agent to the verb) to identify the components of the circuit.

3. Performing a qualitative comparison of the identified circuits, focusing on how the models adapt their mechanisms to handle the distinct typological features of each language (word order vs. postpositions vs. agglutinative suffixes).

4. Construction and comparison of circuits for such features across the languages.

We are on track to complete these steps and present a comprehensive comparative analysis in our final report. The relevant repository for this project is on GitHub (this link).

[2] Precise details are once again in the aforementioned notebook.

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.

Poulami Ghosh, Shikhar Vashishth, Raj Dabre, and Pushpak Bhattacharyya. A Morphology-Based Investigation of Positional Encodings. *Preprint*, arXiv:2404.04530.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer Language Models without Positional Encodings Still Learn Positional Information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390. Association for Computational Linguistics.

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.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. In-context Learning and Induction Heads. *Preprint*, arXiv:2209.11895.

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.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small. *Preprint*, arXiv:2211.00593.