— Research Statement

VSDS Mahesh Akavarapu
✉ maheshak@cse.iitk.ac.in • mahesh-ak.github.io

Computational Historical Linguistics

My research aims to develop and improve methods that automate procedures in historical linguistics by leveraging novel methods inspired by computational biology. These tasks, including the identification of cognates, reconstruction of proto-languages, and classification of languages, traditionally involve significant manual effort. By introducing the Cognate Transformer (Akavarapu and Bhattacharya, 2023a), a model adapted from a protein language model, we have achieved significant improvements in automated phonological reconstruction. This model, especially effective when pre-trained on a masked word prediction task, predicts ancestral proto-words and reflex words in daughter languages with high accuracy. Additionally, we have incorporated elements from protein structure predictors to enhance the Cognate Transformer for automated cognate detection, advocating a supervised approach that outperforms the prior methods with sufficient labeled data and significantly reduces computation time (Akavarapu and Bhattacharya, 2024a). Furthermore, inspired by molecular phylogenetics, we have proposed a novel likelihood ratio test to determine genetic relationships among languages by analyzing phonetically conserved sites (Akavarapu and Bhattacharya, 2024b). This approach addresses the limitations of existing permutation-based tests, reducing false positives, and supports the existence of significant language sub-groupings, such as those within the Macro-Mayan and Nostratic families namely Mayan-Mixe-Zoquean and Indo-European-Dravidian. Through these innovations, my research contributes to the automation and refinement of key processes in historical linguistics, providing tools that offer greater accuracy and efficiency.

In the future, I hope to address the problem of automatic sound correspondence detection along with improving the above methods based on their limitations.

Low Resource Natural-Language Understanding: Sanskrit

I am also focusing on low-resource natural-language understanding, particularly in the context of Sanskrit. We have pre-trained BERT-like models on Sanskrit and evaluated them on a classification benchmark post-fine-tuning, where we found them to perform better than other models such as ELMo, FastText, etc. (Akavarapu and Bhattacharya, 2023b). Fast-forwarding into the age of LLMs, none so far could able to demonstrate conversational abilities in Sanskrit except sufficiently large and multilingual LLMs starting from GPT-4, however, with the problem of hallucination. I am currently working on enhancing the Sanskrit question-answering capabilities of GPT-4 using Retrieval-Augmented Generation (RAG) to avoid hallucination. This ongoing work aims to provide more accurate tools for processing and understanding ancient texts in low-resource languages, contributing to both computational linguistics and the study of classical languages.

Future Research Goals

The end-to-end neural architecture training paradigm of solving problems has now been largely replaced by algorithmic reasoning using LLMs owing to techniques like Chain-of-Thought (with backtracking, etc.). However, such abilities are still limited for **low-resource** languages. I hope to contribute in the future to a few such languages by incorporating linguistic knowledge into the solutions of natural-language understanding problems and the generation of synthetic training data. Likewise, the application of LLMs to the problems of historical linguistics is limited since they are not capable of processing words at the phoneme level owing to limitations in **tokenization**. The same problem would exist for biological sequences. I also hope to work in this direction by exploring the possible ways of circumventing the limitations of tokenization in LLMs in general, for instance, by actually 'reading' the sequences utilizing Large Vision-Language Models.

Publications

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023a. Cognate transformer for automated phonological reconstruction and cognate reflex prediction. In *Proc. of EMNLP 2023*.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2023b. Creation of a digital rig Vedic index (anukramani) for computational linguistic tasks. In *World Sanskrit Conference 2023*.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024a. Automated cognate detection as a supervised link prediction task with cognate transformer. In *Proc. of EACL 2024*.

V.S.D.S.Mahesh Akavarapu and Arnab Bhattacharya. 2024b. A likelihood ratio test of genetic relationship among languages. In *Proc. of NAACL 2024*.