Teaching Machines Sanskrit: Modern NLP for an Ancient Language

KID: 20250318 | Mr Sushant Dave

Sanskrit is one of the oldest living languages in the world. Its earliest known texts, the Vedas, date back over three thousand years. For centuries it served as a vehicle for philosophy, science, literature, and culture across the Indian subcontinent. Despite its systematic grammar and immense intellectual heritage, Sanskrit has remained one of the most difficult languages for computers to process.

Indic language Processing (ILP) group at HST tries to address this challenge: to bring modern Natural Language Processing (NLP) techniques to Sanskrit, bridging ancient knowledge with the latest in machine learning. This article presents the journey — the linguistic challenges, computational experiments, key breakthroughs, and why this work matters for both Sanskrit and the future of AI.

Why Sanskrit Challenges Computers

Unlike English or other European languages, Sanskrit poses unique hurdles for NLP.

Inflectional Richness

Every noun root in Sanskrit can appear in 72 different forms (3 genders × 3 numbers × 8 cases). Verbs are even more complex, with up to 900 forms across tenses, moods, persons, and numbers.

Free Word Order

The sentence "Dog bites man" can be rearranged in six different ways in Sanskrit without changing meaning, because word endings encode grammatical roles. Standard NLP methods, which rely heavily on word order, struggle with this.

Sandhi (Euphonic Combination)

Sanskrit words often merge at boundaries, changing sounds and producing new fused words. For example, guru + upadesha becomes gurupadesha. Splitting these correctly is essential but non-trivial.

Samasa (Compounds)

Beyond Sandhi, Sanskrit allows long compounds where words are joined semantically to form entirely new terms. These compounds can span multiple words and change meaning unpredictably.

Scarcity of Digital Resources

Compared to English or Hindi, Sanskrit lacks large annotated datasets. Most texts exist in traditional print editions, often without digital markup.

Together, these properties make Sanskrit a fascinating but formidable candidate for NLP research.

Why Apply NLP to Sanskrit?

The motivation is both cultural and scientific. Sanskrit texts preserve knowledge in fields ranging from astronomy to medicine. Many remain untranslated or partially studied. Computational tools could:

- Enable faster translation and annotation of manuscripts.
- Support semantic search across massive corpora.
- Assist scholars in identifying linguistic or philosophical patterns.
- Preserve and make accessible the cultural heritage of India.

At the same time, Sanskrit serves as a test case for building NLP tools for other morphologically rich and under-resourced languages worldwide.

Word Vectors for Sanskrit

Word embeddings, or word vectors, represent words as points in a numerical space, capturing semantic similarity. For example, in English embeddings,

vector(King) – vector(Man) + vector(Woman) ≈ vector(Queen)

We trained Word2Vec models on a Sanskrit corpus of over half a million unique tokens, including epics like the Mahabharata and classical works from the Digital Corpus of Sanskrit.

Findings

Verbs clustered meaningfully — many neighbors shared the same root.

Nouns were less coherent — inflectional complexity scattered related forms.

Compositional tests (e.g., gender or number changes) worked only sporadically.

The conclusion was clear: morphological preprocessing is a prerequisite before embeddings can capture Sanskrit semantics effectively.

The conclusion was clear: morphological preprocessing is a prerequisite before embeddings can capture Sanskrit semantics effectively.

Language Modeling

N-gram and Neural Models

We built statistical N-gram models (unigram, bigram, trigram, 4-gram). While English models achieve perplexity scores around 100, Sanskrit models scored in the thousands — showing very poor predictive power. We then trained recurrent neural networks (RNNs) with LSTM cells on the same corpus. Training loss dropped, but validation loss stagnated — a sign of overfitting. The models memorized patterns but failed to generalize.

Inference - Without morphological analyzers to normalize word forms, higher-level tasks like language modeling are doomed to underperform.

Sandhi

Sandhi rules describe how sounds shift at word boundaries. Splitting a fused word (Vichchheda) is context-dependent and often ambiguous. We have worked on expanding the scope of earlier work done by other researchers on Sandhi. One popular approach is formulating Sandhi splitting as a sequence-to-sequence learning problem. Instead of coding all rules by hand, a neural network can be trained to predict splits directly from data. Our group has worked on improving the existing neural approaches with better training data.

Understanding Pratyaya (Suffixes)

Another focus is Pratyaya — suffixes used to derive new words. Two major types are:

- Kridanta (from verbs, e.g., participles and derivatives).
- Taddhita (from nouns/adjectives, forming secondary derivatives).

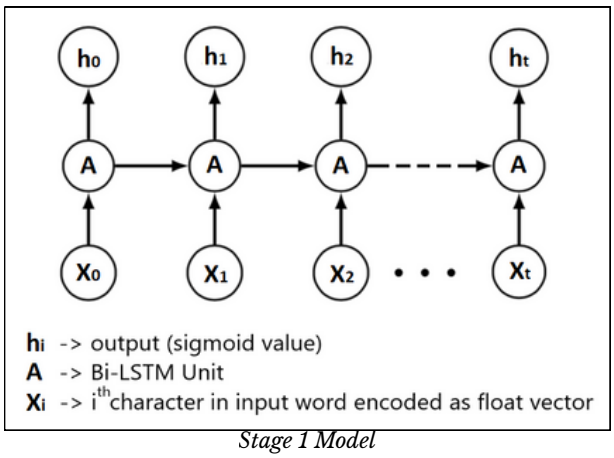
As of now, no exhaustive and reliable benchmark exists for suffix analysis in Sanskrit. ILP team is working on creating PratyayaKosh, the first benchmark corpus for suffix analysis. It includes derivative nouns annotated with their suffixes, enabling systematic evaluation of computational tools. This will be made available to the general public in the near future. Using this dataset, our team is working on Neural models and Finite State Machines that can outperform existing systems. This work can set a new baseline for Pratyaya analysis and open the door for advanced future research.

Handling Samasas

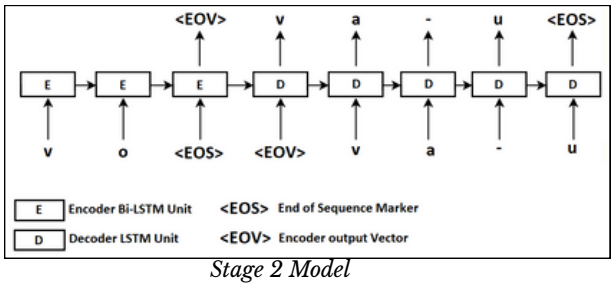
Samasas add an additional layer of complexity to the Sandhi. Samosas are compounds with mandatory Sandhi if possible as per the rules of Panini.

Approach

We took a leaf out of our Sandhi work and tried to figure out split points, before actual split. Only, this time we are looking for multiple split points and not a single one. As before, we have 2 Neural Networks doing the job.



Stage 1 Model



Stage 2 Model

Results

- The neural Samasa model holds its own against Transformer models despite having less than 1% of the trainable parameters.
- Easier to train, fewer data-points and computer resources needed to train
- Again, it required no external lexical resources.

Key Research Outcomes

1. Limits of standard NLP — word embeddings and N-gram models underperform without morphological preprocessing.
2. PratyayaKosh dataset — a standardized benchmark for suffix analysis.
1. Neural compound analysis models — outperformed existing rule-based tools.
2. Emphasis on morphology — highlighted that progress in Sanskrit NLP depends on strong morphological analyzers.

Why This Matters

- Sanskrit NLP is not just an academic exercise. By unlocking Sanskrit texts, we can:
- Make centuries of Indian knowledge in science, medicine, and philosophy accessible.
- Support digital preservation of cultural heritage.
- Build tools for education, research, and even popular use.
- Provide models for how AI can address other low-resource, structurally complex languages.

Future Scope

The journey has just begun. Future work could involve:

- Larger datasets from digitized Sanskrit manuscripts.
- Large language models For all the challenges presented, LLMS are inevitable owing to their generalization power.
- Hybrid methods combining rule-based Paninian grammar with LLMs to make training feasible.
- User-facing applications:
    1. Digital assistants that read and translate Sanskrit.
    2. Search engines for Sanskrit literature.
    3. Educational platforms for students and scholars.

The ultimate dream: AI that can not only read but also converse in Sanskrit, bridging millennia of knowledge with the present.

Sanskrit presents challenges far beyond most modern languages, but these challenges also offer opportunities. By combining ancient linguistic insights with modern AI, we can unlock texts that hold timeless wisdom. Our work at ILP on Sandhi, Pratyaya, Samasa and morphological analysis is only the first step. The road ahead is long, but the vision is clear: a future where machines understand Sanskrit as fluently as humans once did.

“Sanskrit presents challenges far beyond most modern languages, but these challenges also offer opportunities. By combining ancient linguistic insights with modern AI, we can unlock texts that hold timeless wisdom. Our work at ILP on Sandhi, Pratyaya, Samasa and morphological analysis is only the first step”

Mr Sushant Dave
Research Scholar, Dept of HST