Automatic Accent Restoration in Vedic Sanskrit with Neural Language Models

Yuzuki Tsukagoshi & Ikki Ohmukai

Digital Humanities Section, Graduate School of Humanities and Sociology
The University of Tokyo

Background and Problem

- Vedic Sanskrit: oldest layer of Sanskrit.
- **Distinctive pitch accent**: typically one accented syllable per word, crucial for philological and linguistic analysis.
- **Some digital texts lack accent marks** (even UD dataset lacks them!).
- Task: automatically restore accent marks in transliterated Vedic Sanskrit.
- Sequence prediction is challenging:
  - surface form often insufficient,
  - accent depends on phonology, morphology, and syntax.

Contributions

- Construct a large accented Vedic corpus from TITUS.
- Formulate accent restoration as sequence-to-sequence generation.
- Fine-tune three LLMs:
    - LoRA-adapted Llama 3.1 8B Instruct,
    - OpenAI GPT-4.1 nano,
    - Google Gemini 2.5 Flash.
- Evaluate with precision/recall/F1, CER, WER, and ChrF1.
- Show that fine-tuned models substantially outperform untuned baselines.

Accent System Overview

- One accent per word in principle; encoded with acute (*á*) and grave (*à*).
  - Exceptions: enclitics, finite verbs in main clauses, vocatives, etc.
- **Accent is not purely lexical**: inflection can shift accent position; analogical change adds irregularity.
  - Example (√*as*): *s-án* (nom. sg., suffix accent) vs. *s-at-ás* (gen. sg., ending accent).
  - Shift patterns (e.g., acro-/protero-/amphi-/hysterodynamic)
- **Accents in Compounds**: endocentric (final-member accent) vs. exocentric (first-member accent).

Dataset

- **Source**: **TITUS**; Saṃhitā (poetry) and Brāhmaṇa (prose) texts were extracted and segmented into lines (*paada*) or sentences.
- **ISO 15919 transliteration** with accent marks (acute/grave = *udātta*/*svarita*) on vowels.
- **Supervision**: input without accents, output with original accents.
- **Size**: **108,076** samples, avg 6.03 words, 133,873 unique forms.
- Split: 8:1:1 train/validation/test.

Models

- Open-weight model:
    - Llama 3.1 8B Instruct (LoRA fine-tuning).
- Proprietary models:
    - OpenAI GPT-4.1 nano,
    - Google Gemini 2.5 Flash.
- All models trained in seq2seq style:
    - input: unaccented Vedic text,
    - output: same text with restored accent marks.

Evaluation Setup

- Test set: held-out accented sentences from all texts.
- Primary metrics (on vowels):
    - precision, recall, F1 for accent placement.
- Additional metrics:
    - character error rate (CER),
    - word error rate (WER),
    - ChrF1.
- Compare fine-tuned models with untuned baselines and compare the performance of each model.

Results

Model	Precision↑	Recall↑	F1↑	CER↓	WER↓	ChrF1↑
GPT-4.1 nano (Before SFT)	0.609	0.020	0.039	0.288	0.858	45.6
GPT-4.1 nano (After SFT)	0.752	0.676	0.712	**0.062**	0.322	79.6
Gemini 2.5 Flash (Before SFT)	0.551	0.191	0.284	0.698	0.863	22.6
Gemini 2.5 Flash (After SFT)	0.789	0.771	0.780	0.109	0.249	83.5
Llama 3.1 8B (Before SFT)	0.452	0.034	0.064	0.249	0.894	48.1
Llama 3.1 8B (After SFT)	**0.916**	**0.841**	**0.877**	0.096	**0.161**	**87.5**

Bold indicates the best value per metric.

Error Analysis & Examples

- Frequent errors: misplaced accents in paradigms; compound interpretation ambiguities; rare lexemes.
- Many subtle alternations are captured; models leverage morpho-syntactic and phonological cues.
- Examples: correct common inflectional patterns; occasional errors on irregular/rare forms; mostly correct compound accents with edge-case mistakes.

Conclusion & Next Steps

- First LLM-based approach to Vedic accent restoration with strong results.
- Released large accented corpus
- Next: integrate into broader Vedic NLP (sandhi, parsing, MT, etc.); explore joint modeling.

Dataset on Hugging Face

Questions on Slido