Accent Placement Models for Rigvedic Sanskrit Text

Akhil Rajeev P and Annarao Kulkarni

C-DAC Bangalore

Background

Rigvedic Sanskrit employs a **phonologically contrastive pitch-accent system** that is fundamental to its oral transmission, interpretation, and prosodic structure. Each syllable may bear one of three tonal categories: *udātta* (raised pitch, typically unmarked), *anudātta* (lowered pitch, marked below the character; U+0952), and *svarita* (falling pitch, marked above the character; U+0951). Unlike Classical Sanskrit, these accents are linguistically meaningful and directly influence recitation and metrical realization.

In modern digital corpora, however, accent marks are frequently **omitted or inconsistently encoded** due to limitations of early text standards, OCR pipelines, and poor support for Unicode combining diacritics. Consequently, much of the available Rigvedic text exists in an **accent-stripped form**, resulting in the loss of crucial phonological and prosodic information that is essential for linguistic analysis and faithful oral rendering.

Automatic accent restoration poses a challenging sequence prediction problem, as accent placement depends on morphology, phonology, sandhi, and metrical constraints, while being realized as **sub-character combining symbols**. This makes standard metrics such as word error rate insufficient for isolating tonal errors, motivating byte-level modeling and diacritic-focused evaluation for accurate accent prediction.

Dataset

We use an **in-house, validated Rigveda corpus** developed at C-DAC, comprising **10,552 hymns** organized across **10 maṇḍalas** and **1,028 sūktas**. From this resource, we construct a **parallel corpus of 22,740 aligned verse pairs**, where each entry pairs an **unaccented verse** with its **diacritically marked (accented) counterpart** for supervised training and evaluation.

अग्निमीळे पुरोहितं यज्ञस्य देवमृत्विजम्।

Figure 1. Unaccented Rigveda Sentence.

अ॒ग्निमी॑ळे पु॒रोहि॑तं य॒ज्ञस्य॑ दे॒वमृ॒त्विज॑म्।

Figure 2. Accented Rigveda Sentence.

The C-DAC Rigveda parallel corpus will be made available through the Indian Knowledge Base platform.Ancient Indian Heritage Knowledgebase (AIHKB) portal encompasses Ancient Indian Scientific Heritage text corpus, created by IHLC team over the last two and half decades.A beta version of the portal is released alongside this paper.

Problem Statement

Despite the linguistic importance of pitch accents in Rigvedic Sanskrit, most available digital texts lack accentual information due to encoding and OCR limitations. The task is to automatically restore correct pitch accents in unaccented Rigvedic verses by accurately predicting Unicode combining diacritics while preserving the underlying grapheme sequence.

Proposed Approach

We study accent restoration as a byte-level sequence generation task and evaluate **three complementary modeling approaches** on the aligned unaccented–accented Rigvedic corpus, covering both full and parameter-efficient fine-tuning regimes.

(1) Full Fine-tuning of Sanskrit ByT5 We fine-tune **Sanskrit ByT5** (Nehrdich et al.), a byte-level encoder–decoder Transformer pretrained on Sanskrit text, by updating all model parameters. Byte-level modeling enables direct handling of Unicode combining diacritics used for Vedic pitch accents.

(2) LoRA-based Fine-tuning of Sanskrit ByT5: To enable parameter-efficient adaptation, we integrate **LoRA adapters** into the attention projections of Sanskrit ByT5 while freezing the base model weights. This approach significantly reduces the number of trainable parameters while preserving accent restoration capability.

(3) BiLSTM–CRF Sequence Tagging Baseline: As a non-Transformer baseline, we implement a character-level BiLSTM–CRF model that predicts accent labels for each character. This setup provides a comparison against traditional sequence labeling approaches commonly used in low-resource linguistic tasks.

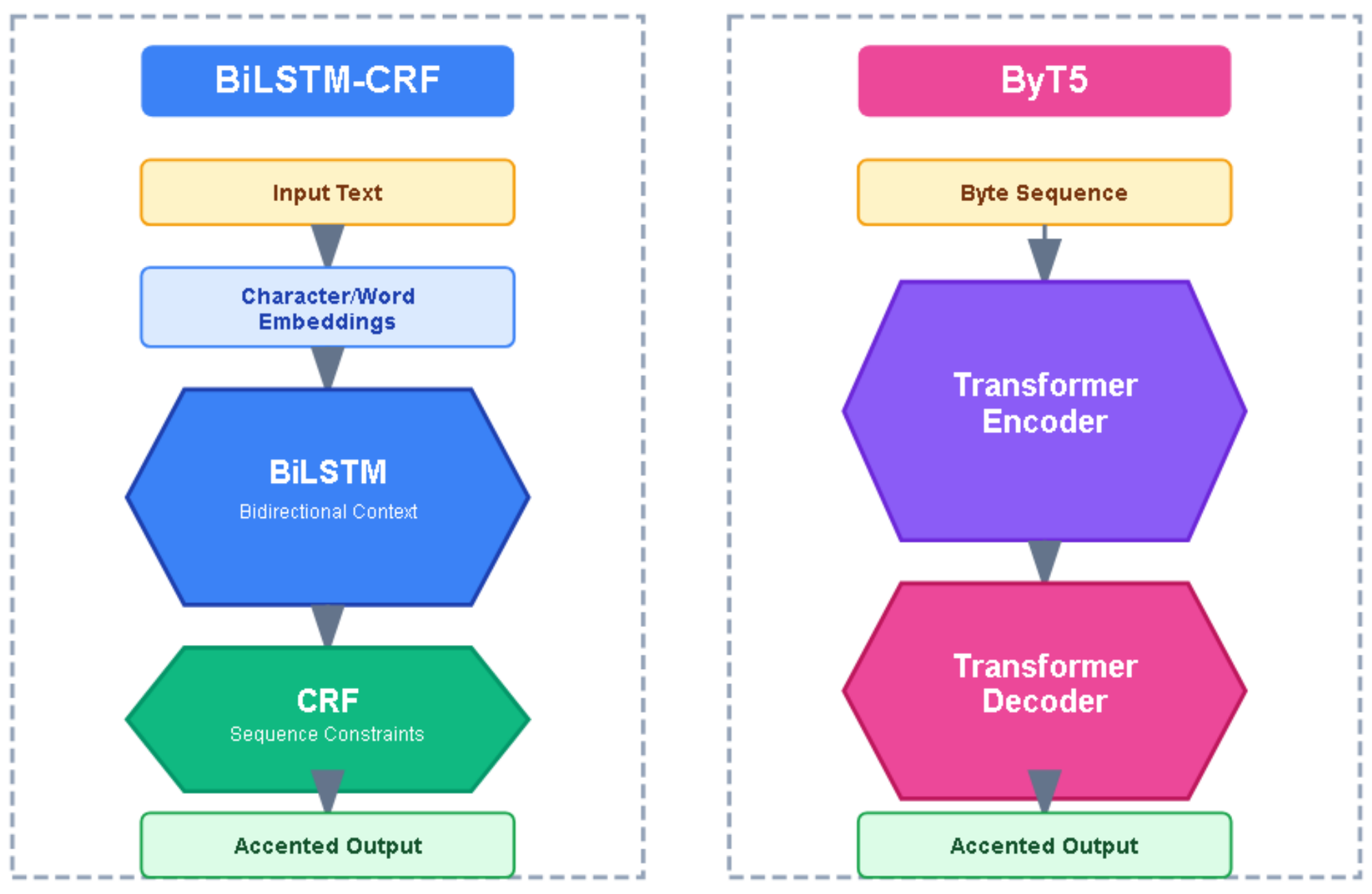

Figure 3. ByT5 vs BILSTM-CRF

The BiLSTM–CRF baseline performs character-level accent tagging and is lightweight, but lacks pretraining and long-range context, leading to high diacritic error rates. Full fine-tuning of Sanskrit ByT5 models accent restoration as byte-level generation and achieves the best accuracy at the cost of high computation while LoRA-based fine-tuning offers a PEFT alternative which offers a middle ground .

Evaluation Metrics

System performance is evaluated using **three complementary string-level metrics** that jointly capture lexical accuracy, orthographic fidelity, and accent-specific correctness.

- **Word Error Rate (WER)** measures **token-level edit distance** (insertions, deletions, substitutions), reflecting overall lexical correctness.
- **Character Error Rate (CER)** computes **character-level edit distance** (excluding whitespace), capturing fine-grained orthographic deviations.
- **Diacritic Error Rate (DER)** isolates errors in **accent diacritics** alone, ignoring base characters, and directly measures tonal accuracy.

Together, these metrics provide a layered view of model behavior: WER reflects global token accuracy, CER captures character integrity, and DER specifically evaluates accent restoration at the sub-character level.

Results

Method	WER	CER	DER
Full Fine-tuning (Sanskrit ByT5)	**0.1023**	**0.0246**	**0.0685**
BiLSTM–CRF	0.2367	0.0448	0.3197
LoRA Fine-tuning (Sanskrit ByT5)	0.3614	0.1042	0.1598

Table 1. Performance of three modeling approaches on Rigvedic accent restoration. Best scores are highlighted in bold.

For Further Information

(a) Paper Link

(b) C-DAC Indian Knowledgebase