Lexical Accent Prediction in Ṛgvedic Sanskrit Using Morphological Information

Anustup Bhattacharyya
HSS, IIT Bombay
bhattacharyya@iitb.ac.in

Akshay Gaikwad
CSE, IIT Bombay
akshai@cse.iitb.ac.in

Malhar Kulkarni
HSS, IIT Bombay
malharku@gmail.com

Rushikesh K. Joshi
CSE, IIT Bombay
rkj@cse.iitb.ac.in

Abstract

Lexical pitch accent in Ṛgvedic Sanskrit is phonemic and strongly conditioned by morphology, yet automatic accent prediction has received little attention. We address this problem using the VedaWeb corpus of the Ṛgveda, formulating accent prediction as a character-level sequence labeling task over unaccented word forms with morphological tags. We apply bidirectional LSTM and Transformer models, with and without morphological feature embeddings. Our results show that morphological information is essential in aiding sequential models for accurate accent placement. Morphological features such as case, gender, number, etc. are found to significantly improve performance. The BiLSTM model with morphological features is found to achieve the best overall results.

1 Introduction

Vedic Sanskrit is a pitch accent language. Pitch is phonemic in nature, i.e., it brings about a change in meaning and is described to be musical in nature (Macdonell(1993)). Thus, a change in the pitch-accent of a vowel/syllable in a word would also mean a change in the meaning of the word, for example: ápas (neuter) : 'work' vs. apás- adj. 'active' or bráhman- n. 'poetic formula' vs. brahmán- m. 'poet, priest'. Vedic pitch accent also corresponds to the pitch accent of its sister Proto Indo-European languages, namely Greek and Latin, where some texts still preserve accent. In the later stages, the pitch accent is lost giving rise to a stress accent (a stress accent language stresses on a particular syllable in a word, like English) in the classical stage.

Accent, although a suprasegmental feature and is strongly conditioned by morphology (e.g., root/stem and its inflectional class/ending), on which Renou (Renou(1952)) comments: "En fait, l'uda̎tta coïncide en principe avec le degré plein des formations alternantes: sa présence ou son absence est une donnée morphologique."

1.1 Description of the accentual system

Mainly, three pitches are differentiated: the acute (*udātta*), the grave (*anudātta*) and the circumflex (*svarita*), which is described as a mix of the grave and the acute. In the R̥gveda, the circumflex rises to a higher pitch than the acute accent before falling down, which according to Macdonell (Macdonell(1993), p.449) might be the reason why the acute remains unmarked and the other two are committed to writing in the *Devanāgari* script, where the grave is indicated by a horizontal line below the vowel (vowel-consonant combination) and the circumflex by a vertical bar on top of the vowel (or the vowel-consonant glyph). In Roman with diacritics, the (*udātta*) or the principal accent is marked by an acute accent mark (´). Here is the same *Rk* written in *devanāgari* and in roman with diacritics along with accent marks:

agním i̎ḷe puróhitaṁ
yajñasya devám ṛtvíjam
hóta̎raṁ ratnadhátamam

अ॒ग्निमी॑ळे पु॒रोहि॑तं
य॒ज्ञस्य॑ दे॒वमृ॒त्विज॑म्
होता॑रं रत्न॒धात॑मम्

While the main accent is the *udātta*, the *udātta* and the svarita are enclitic accents used before and after an *udātta*, respectively. Instances of the svarita are found to be present where they occur independently when an *udātta* gets deleted. The accentual properties of Vedic Sanskrit are also present in the sister Indo-European languages of Latin and Greek. A word can have only one udatta, i.e., a main accent. Some words can be bereft of an acute accent, i.e., finite verbs in the principal clause, particles, enclitic pronouns, vocatives etc.

1.2 Indian Grammatical Treatises on Accent

The rules dealing with Vedic or later accents form a separate topic of study, often less-studied than other grammatical topics in modern times. The Aṣṭa̎dhya̎yi̎ has dealt extensively with accents. Other texts as uṇa̎di su̎tras, phiṭ su̎tras and the pra̎tiśa̎khyas have dealt with accents. The Vaiya̎karaṇa Siddha̎nta Kaumudi̎, a thematic rearrangement of the Aṣṭa̎dhya̎yi̎ has a separate chapter dedicated to the accentual rules, the Svaraprakaraṇa. We also find singular treatises dealing only with the accentual rules like the Svara̎ṇkuśa of Jayantasva̎min with Nilakaṇṭha's commentary, Svaralakṣaṇa (authorship unknown), Svarasiddha̎ntamañjari̎ of Nr̥siṁhapaṇḍita, Svarasiddha̎ntaandrika̎, Svaraprakriya̎ of Viṭṭhaleśa, Svaraprakri̎ya̎ of Bhaṭṭoji Di̎kṣita, Svaraprakri̎ya̎ of Ra̎macandraśeṣa (Abhyankar(1916)) are other known mostly unpublished on the same subject. The third chapter of the R̥gvedaprātiśākhya (Deva Shastri(1937)) deals briefly with the accentual patterns in the R̥gveda.

Although these accents are not exclusively mentioned for Vedic, but there is some relation between the vedic accent and to that of the later accentual system, on which Abhyankar (Abhyankar(1916)) comments: "...all those (svara) rules except a few that are specified for vedic literature, apply to the vedic literature as well as to the spoken language... are applicable universally."

The navya-vya̎karaṇa texts eschew the discussion on accent along with special rules for Vedic, possibly for reasons of brevity and simplicity. This also points to the fact that pitch accent is not only lost but also its relevance to the grammarians as a separate topic of study is lost over time.

1.3 Related Work

The problem of modeling and restoring lexical accent in Vedic Sanskrit has traditionally been addressed within descriptive and philological scholarship, where accent placement is treated as a phonemic phenomenon closely tied to morphology and grammatical function. However, computational approaches to accent prediction have remained limited until recently, largely due to the scarcity of digitally available accented corpora.

Recent work has begun to explore neural approaches to automatic accent restoration in Vedic Sanskrit. Tsukagoshi and Ohmukai (Tsukagoshi and Ohmukai(2025)) apply large neural language models to the task of restoring accent marks in transliterated Vedic text, demonstrating that pretrained models can capture accentual regularities when fine-tuned on parallel accented–unaccented data.

Complementing this, Rajeev and Kulkarni (Rajeev and Kulkarni(2025)) construct a parallel corpus of accented and unaccented Rigvedic verses and evaluate both recurrent and Transformer-based models for accent placement, showing that lexical sequence modeling substantially outperforms context-free baselines.

These studies establish the feasibility of data-driven accent prediction and highlight the importance of sequential context, where they feed words in sequence, which also captures syntactic relations. Building on this line of work, our study differs in explicitly incorporating morphological feature embeddings into character-level sequence models, allowing us to directly test the hypothesis, well motivated in traditional grammar, that grammatical categories such as case, gender, and number play a crucial role in determining lexical accent placement in R̥gvedic Sanskrit. Thus, the focus of this paper is to empirically establish the importance of morphological features on lexical accent placement through AI techniques.

2 Literature Review

2.1 Dataset: the VedaWeb, a morphologically-tagged Corpus of the R̥gveda

VedaWeb
Online Research Platform
for Old Indic Texts

"The VedaWeb platform provides access to ancient Indian texts written in Vedic Sanskrit, which are enriched with morphological and metrical annotations, making them searchable according to lexicographical and corpus-linguistic criteria" Kölligan et al. (Kölligan et al.(2021)Kölligan, Neuefeind, Reinöhl, Sahle, Casaretto, Bunselmeier, Coenen, Fischer, Kiss, Korobzow, Rolshoven, Halfmann, and Mondaca).

The development of the VedaWeb platform has significantly changed this landscape by providing richly annotated Vedic texts with detailed morphological information, enabling systematic corpus-based and computational investigations of accentuation patterns. The version of the corpus used is in .xlsx format. [1]

2.1.1 Morphological Glosses

The morphological annotation follows the Leipzig Glossing Rules (Kölligan et al.(2019)Kölligan, Neuefeind, Kiss, Mondaca, Reinöhl, and Sahle). These glosses are 38 in number (3 for person, 8 for cases, 3 for voice, 7 for tense, 3 for number, 3 for gender, 5 for mood, and 4 for converb, and 2 for participles. The abbreviations are listed below in table 1

Code	Description	Code	Description	Code	Description
1	first person	FUT	future	NOM	nominative
2	second person	GEN	genitive	OPT	optative
3	third person	IMP	imperative	PASS	passive voice
ABL	ablative	IND	indicative	PL	plural
ACC	accusative	INF	infinitive	PLUPRF	past perfect
ACT	active	INJ	injunctive	PPP	na participle perfective passive
AOR	aorist	INS	instrumental	PPP	ta participle perfective passive
COND	conditional	IPRF	imperfect	PRF	perfect
CVB	converb	LOC	locative	PRS	present
DAT	dative	M	masculine	PTCP	participle
DU	dual	MED	middle voice	SBJV	subjunctive
F	feminine	N	neuter	SG	singular
				VOC	vocative

Table 1: Moptohlogical glosses in the Vedaweb corpus

2.1.2 Corpus Description

The corpus has 32 columns out of which the initial five columns indicate the verse, stanza, and the remaining columns indicate the grammatical categories and their wider intricacies, the type of lemma it belongs to and the meaning of the words. Each row thus represents a word and the initial columns provide a reference for its actual location in the R̥gveda. The column 'morphological tag' provides a composite grammatical tag and the following columns provide the other possible tags for the same word. We investigate how morphological features of a word influence it's accent placement.

[1] https://github.com/cceh/c-salt_vedaweb_sources/blob/master/rigveda/versions/zurich.xlsx

The following morphological categories available as values of the columns containing grammatical categories in the corpus are used for the purpose of our training task:

- Case: Nominative, Accusative, Instrumental, Dative, Ablative, Genitive, Locative, Vocative
- Number: Singular, Dual, Plural
- Gender: Masculine, Feminine, Neuter
- Person: 1st, 2nd, 3rd
- Tense/Aspect: Present, Aorist, Future, Perfect, Pluperfect, Imperfect, Conditional
- Mood: Indicative, Subjunctive, Imperative, Injunctive, Optative
- Voice: Active, Middle, Passive
- Other tags: Converb, Infinitive, ta-na participles

2.2 Problem Statement

The problem is to find out whether given an unaccented word and its morphological tags we can predict better the lexical accent of the word, as compared to predicting it without using the morphological features. One way to solve the problem is to create a rule-based system, where the accentual rules in several grammatical texts could be modeled. Such a model would also require the addition of all the other rules of inflection and derivation into the same system to produce the correct accented word form. Given the voluminous nature of this method, it would be challenging to model the accentual properties along with other rules into a rule-based system, which again would give rise to rule conflicts due to the scope of the rules and their application.

A viable solution is to use a Deep Learning model with an accented word along with its grammatical features, so that given an unaccented word with its morphological tags, the model can predict the accented word-form.

3 Methodology

3.1 Architecture Overview

The proposed framework integrates character-level representations with handcrafted morphological features for word-level classification. The architecture consists of three stages: input representation, feature fusion, and sequence encoding followed by accent prediction. Two encoder variants are implemented: Bidirectional Long Short-Term Memory (BiLSTM) and Transformer—while maintaining identical input processing and feature fusion mechanisms. This ensures that performance differences arise solely from the sequence modeling component. Both models are applied with and without features to bring out the utility of morphological features in accent prediction. The architecture is shown in Figure 1.

3.2 Input Representation

3.2.1 Character-level Embedding

Each word is decomposed into a sequence of characters. The character sequence is integer-encoded and padded to a fixed maximum length $L = 32$, to accommodate the maximum length of a word in the corpus. A trainable embedding layer maps each character index to a dense vector of dimension $d_c = 64$. This produces a character embedding matrix for a given word w:

$$\mathrm{E}_w \in \mathbb{R}^{L \times d_c}$$

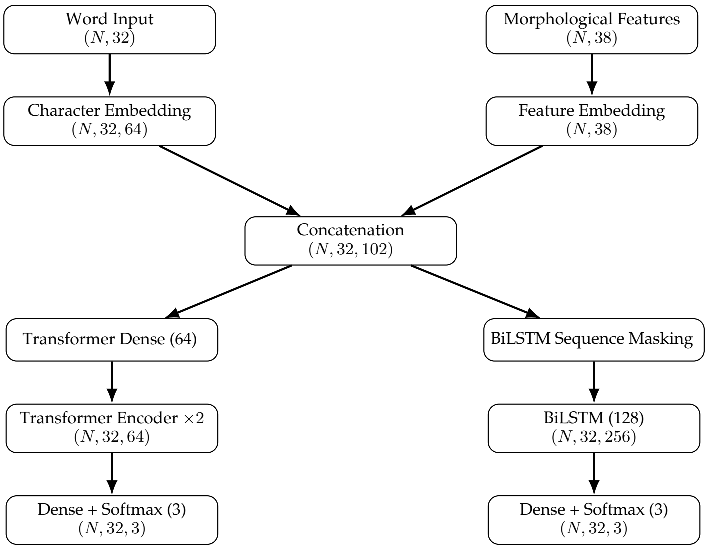

Figure 1: Overall architecture of the proposed character-level accent prediction model

3.2.2 Morphological Feature Embedding

In addition to character-level encoding, each word is represented using structured morphological and linguistic attributes. The feature set includes: case, gender, number, lemma, lemma type, additional noun features, person, mood, tense, voice, and additional verb features.

Each categorical feature is embedded independently using a trainable embedding layer with manually assigned embedding dimensions reflecting relative representational requirements. The embedding dimensionalities are allocated as follows: case (4), gender (2), number (2), lemma (8), lemma type (4), additional noun features (4), person (2), mood (3), tense (3), voice (2), and additional verb features (4). The concatenation of these feature embeddings results in a unified morphological representation of dimension $d_f = 38$.

Since morphological attributes are defined at the word level while character embeddings operate at the sequence level, the resulting 38-dimensional feature vector is repeated across all character positions to obtain:

$$\mathrm{E}_f \in \mathbb{R}^{L \times d_f}$$

This enables alignment between morphological information and character-level representations.

3.3 Feature Fusion

The character embedding matrix and the repeated morphological feature matrix are concatenated along the feature dimension to form a unified representation:

$$\mathrm{X} = [\mathrm{E}_c; \mathrm{E}_f]$$

The resulting fused representation has dimensionality:

$$\mathrm{X} \in \mathbb{R}^{L \times (d_c + d_f)}$$

In the current implementation, $d_c = 64$ and $d_f = 38$, resulting in a 102-dimensional feature vector per character position.

Dense(64) layer is applied after concatenation for ; the fused representation is directly provided to the sequence encoder.

3.4 Sequence Encoding

The fused sequence representation is processed using two encoder architectures: a Bidirectional Long Short-Term Memory network and a Transformer encoder. Both encoders operate on the same 102-dimensional input representation.

3.4.1 Bidirectional LSTM Encoder

The fused input representation X $\in \mathbb{R}^{32\times 102}$ is processed using a Bidirectional LSTM layer with 128 units in each direction, producing a 256-dimensional contextual representation per timestep. A masking layer is applied prior to the LSTM to ignore padded positions. Design parameters:

- Character embedding dimension: 64
- Morphological feature dimension: 38
- LSTM units: 128 (per direction)
- Output hidden dimension: 256
- Dropout rate: 0.3
- Batch size: 64
- Epochs: 10

3.4.2 Transformer Encoder

The fused representation is processed using stacked Transformer encoder blocks operating with a fixed model dimension $d_{model} = 64$. Each block consists of multi-head self-attention followed by a position-wise feed-forward network, with residual connections and layer normalization. Design parameters:

- Character embedding dimension: 64
- Number of attention heads: 4
- Feed-forward dimension: 128
- Number of Transformer blocks: 2
- Dropout rate: 0.3
- Batch size: 64
- Epochs: 10

3.5 Classification Layer

The final encoder output has dimensionality

$$\mathbb{R}^{32\times 64}.$$

After applying dropout (rate = 0.3), a position-wise dense layer with 3 output units (0: 'No accent', 1: 'Acute', 2:'grave') is applied. This produces an output tensor of shape:

$$\mathbb{R}^{32\times 3},$$

corresponding to class probabilities over three target categories for each sequence position. The model is trained using categorical cross-entropy loss with the Adam optimizer.

4 Results, Discussion, and Conclusion

4.1 Results

Table 2 shows the character-level performance of the BiLSTM encoder (character+feature embedding) evaluated on the character-accent classification task. Metrics reported include character-level accuracy, precision, recall, and F1-score for each of the three lables along with aggregated macro and weighted scores of the same.

Class	Precision	Recall	F1-score	Testing size
0 (No Accent)	0.98	0.94	0.96	46684
1 (Acute)	0.61	0.82	0.70	5589
2 (Grave)	0.40	0.87	0.54	108
Macro Avg	0.66	0.88	0.73	52381
Weighted Avg	0.94	0.92	0.93	52381

Table 2: Character-level classification results for BiLSTM with morph. features (80:20 split)

Table 3 summarizes the character-level performance of all models across 80:20 dataset split. We can clearly observe performance improvement by inclusion of morphological features.

Models	Macro Average			Weighted Average		
	Precision	Recall	F1-score	Precision	Recall	F1-score
BiLSTM	0.49	0.89	0.53	0.92	0.78	0.82
BiLSTM + Features	0.66	0.91	0.74	0.94	0.93	0.93
Transformer	0.43	0.82	0.42	0.92	0.68	0.75
Transformer + Features	0.49	0.87	0.52	0.92	0.83	0.86

Table 3: Comparison of character-level performance across models (80:20 split)

Table 4 shows word-level accuracy of the two models with and without the use of morphological features for the task of accent placement prediction. We mark any word as correctly predicted when It can be seen that the number of words which are fully correctly accented is significantly higher when morphological features are used.

Model	Word-level Accuracy	Δ Improvement
BiLSTM	4.05%	–
BiLSTM + Features	60.25%	+56.20
Transformer	0.61%	–
Transformer + Features	29.82%	+29.21

Table 4: Word-level accuracy comparison across models (80:20 split)

4.2 Discussion

These results suggest several key insights:

- Morphological features are crucial: Including grammatical information such as case, tense, gender, and number improves both precision and F1-score for both the models, highlighting the value of the morphological context. It can be noted that without using the knowledge of morphology, the performance of both the models was found to be significantly lower, at both character and word level.

- Transformer models require more data: The under-performance of Transformers as compared to BiLSTM may be due to limited training data. Transformers rely on attention over the full sequence, which can lead to overfitting or under-training when the dataset is small.

- Precision vs Recall tradeoff: The models achieve higher recall as compared to precision. This indicates that the correctly predicting an accented character is higher. However, the models are adding more than one accent mark in a word, when typically there is only one accent in a word. Integrating this method with a few known rules of accents may improve the performance.

- Syntactic context: Syntactic context might aide transformer and other sequential models in showing how the syntactic position of words influence accent placement in a sequence. Whereas the present works taken into account individual words, the influence of word sequencing in sentences may carry more information for accent tagging.

These observations indicate that for low-resource tasks like Sanskrit accent prediction, traditional sequential models like BiLSTM with feature augmentation can outperform more complex architectures such as Transformers. More importantly, morphological information impact the results significantly.

4.3 Conclusion

In this study, we presented a word-level approach for automatic accent detection in Sanskrit. We evaluated BiLSTM, and Transformer models, both with and without morphological features.

The BiLSTM with morphological features emerged as the most effective model, demonstrating the importance of both sequential modeling and morphological features. Transformer-based models underperformed as compared to BiLSTM, highlighting that model complexity must be carefully matched to dataset size and task characteristics.

Future work can explore:

- Data augmentation or larger corpora to improve Transformer performance

- More sophisticated feature embeddings, including phonetic or prosodic features

- Multi-task learning frameworks that jointly predict accent and other linguistic annotations.

- Feed sentence-level embedding of words to DL models, in order to analyze higher-level hierarchal influence on accents.

- Analyzing the weight of morphological features in predicting accent placement.

- Integration of common rules of accentual combinations to improve prediction.

Overall, this study confirms that the use of morphology emerges to be more effective for low-resource linguistic accent labeling tasks such as this one, providing a robust foundation for further research in on accentual prediction tasks.

References

[Abhyankar(1916)] K. V. Abhyankar. Svaraprakriya̎, volume 138 of A̎nanda̎śrama Saṃskṛta Grantha̎vali. A̎nanda̎śrama, Puṇe, 1916. Title page shows S̎ālivāhana-S̎aka 1838 (= 1916 CE).

[Deva Shastri(1937)] M. Deva Shastri. The Ṛgveda Pra̎tiśa̎khya, volume 3. Motilal Banarsidass, Delhi, 1937.

[Kölligan et al.(2019)Kölligan, Neuefeind, Kiss, Mondaca, Reinöhl, and Sahle] D. Kölligan, C. Neuefeind, B. Kiss, F. Mondaca, U. Reinöhl, and P. Sahle. Vedaweb – on the role of annotations in analyzing ancient indo-aryan texts. In Proceedings of the Historical Corpora and Variation Conference (HiCoV), Cagliari, Italy, 2019. URL `https://vedaweb.uni-koeln.de/`.

[Kölligan et al.(2021)Kölligan, Neuefeind, Reinöhl, Sahle, Casaretto, Bunselmeier, Coenen, Fischer, Kiss, Korobzow, Rols
D. Kölligan, C. Neuefeind, U. Reinöhl, P. Sahle, A. Casaretto, J. Bunselmeier, P. Coenen, A. Fischer, B. Kiss, N. Korobzow, J. Rolshoven, J. Halfmann, and F. Mondaca. Vedaweb: Online research platform for old indic texts. `https://vedaweb.uni-koeln.de`, 2021. Accessed: <date of access>.

[Macdonell(1993)] A. A. Macdonell. A Vedic Grammar for Students. Motilal Banarsidass Publ., 1993. ISBN 978-81-208-1052-5. Google-Books-ID: YKI3TQvbsDcC.

[Rajeev and Kulkarni(2025)] P. Rajeev and A. Kulkarni. Accent placement models for rigvedic Sanskrit text. In A. Bhattacharya, P. Goyal, S. Ghosh, and K. Ghosh, editors, Proceedings of the 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA 2025), pages 122–126, Mumbai, India, Dec. 2025. Association for Computational Linguistics. ISBN 979-8-89176-313-5. URL `https://aclanthology.org/2025.bhasha-1.11/`.

[Renou(1952)] L. Renou. Grammaire de la langue védique, volume 9 of Les langues du monde. IAC, Lyon, 1952.

[Tsukagoshi and Ohmukai(2025)] S. Tsukagoshi and I. Ohmukai. Automatic accent restoration in vedic sanskrit with neural language models. In Proceedings of the Bhasha Research Workshop, ACL Anthology, 2025. URL `https://aclanthology.org/2025.bhasha-1.7/`.