Evaluating Transliteration Ambiguity in Adhoc Romanized Sinhala: A Dataset for Transliteration Disambiguation

Sandun Sameera Perera
Informatics Institute of Technology
57, Ramakrishna Road, Colombo 06
Sri Lanka
sameeraperera827@gmail.com

Deshan Sumanathilaka
School of Computing
Swansea University, Swansea
United Kingdom
deshankoshala@gmail.com

Abstract

This paper introduces the first Transliteration disambiguation (TD) dataset for Romanized Sinhala, informally known as Singlish, developed to address the challenge of transliteration ambiguity in backwards transliteration tasks. The dataset covers 22 ambiguous Romanized Sinhala words, each mapping to two distinct Sinhala meanings, and provides 30 Romanized sentences per word: ten for each meaning individually and ten containing both meanings in context. Sentences were initially collected through web scraping and later post-processed using the Claude language model, which offers strong support for Sinhala, alongside a rule-based Romanization process to ensure linguistic quality and consistency. To demonstrate its applicability, the dataset was used to evaluate four existing back-transliteration systems, highlighting their performance in resolving context-sensitive ambiguities. Baseline evaluations confirm the dataset's effectiveness in assessing transliteration systems' ability to handle transliteration ambiguity, offering a valuable resource for advancing TD and transliteration research for Sinhala.

1 Introduction

Romanized Sinhala, often referred to as Singlish, is a widely used informal writing style among Sinhala speakers, particularly in digital communication platforms such as social media, messaging apps, and online forums (De Silva, 2019). In this form, Sinhala words are written using the Latin script, typically without standardized transliteration rules (Liwera and Ranathunga, 2020). As a result, Romanized Sinhala exhibits high variability and ambiguity, posing unique challenges for natural language processing (NLP) tasks such as machine translation, language modeling, and especially backward transliteration, the process of converting Romanized text back into native Sinhala script (Athukorala and Sumanathilaka, 2024).

One of the major challenges in ad-hoc backwards transliteration of Romanized Sinhala is transliteration ambiguity, where a single Romanized word may correspond to multiple Sinhala words with distinct meanings. These ambiguities are often context-dependent and are further complicated by the ad-hoc nature of Romanization used by Sinhala keyboard users (Perera and Sumanathilaka, 2025). Despite the significance of this problem, there has been no dedicated resource to support or evaluate systems designed to resolve such ambiguities through contextual understanding.

To address this gap, we introduce the first transliteration Disambiguation dataset for Romanized Sinhala. This dataset is specifically constructed to evaluate the ability of backward transliteration systems to resolve semantic ambiguities. It focuses on 22 Romanized Sinhala words that each map to two semantically distinct Sinhala words. For each ambiguous word, the dataset provides 30 Romanized sentences: 10 exemplifying one meaning, 10 exemplifying the alternative meaning, and 10 containing both meanings within the same sentence context. Each sentence is paired with its corresponding version in the native Sinhala script.

In addition to constructing the dataset, we evaluate its utility by benchmarking four existing back-transliteration systems using this resource. The evaluation highlights how effectively these systems handle context-sensitive transliteration ambiguities in Romanized Sinhala text. The disambiguation process used by transliteration algorithms for the word 'nthi' is presented in Figure 1.

The contributions of this work are as follows:

- Introducing the first TD dataset tailored for Romanized Sinhala, targeting sentence-level disambiguation of transliterated words.
- Describing a systematic approach to selecting ambiguous words and generating high-quality,

Proceedings of Recent Advances in Natural Language Processing,pages 934–942
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-107

context-rich sentence pairs.

- Benchmarking four existing backwards transliteration systems using the dataset, establishing baseline performance metrics.
- Publicly releasing the dataset to encourage further research in Sinhala NLP and transliteration.

In the subsequent sections, we discuss related literature, detail our approach to dataset construction, outline the evaluation framework, and conclude the paper.

2 Related Work

Transliteration ambiguity and back-transliteration have both been long-standing challenges in Natural Language Processing, particularly for low-resource languages and informal language variants like Romanized Sinhala (Singlish). While several back-transliteration systems have been developed to handle ad-hoc Romanized Sinhala input, the lack of dedicated resources to evaluate TD in such contexts has limited progress in context-sensitive transliteration.

Early work by Sumanathilaka et al. (2023) introduced a hybrid back-transliteration system that combined trigram-based statistical modelling with rule-based techniques. Their model demonstrated moderate success in handling informal and shorthand Romanized Sinhala input and was able to resolve some transliteration ambiguities due to the nature of the n-gram model. However, it was not specifically designed to perform context-aware semantic disambiguation.

More recent advances have leveraged neural language models for more accurate transliteration. Perera et al. (2025) proposed a BERT-based back-transliteration system that incorporates dictionary-based mappings, rule-based processing for out-of-vocabulary cases, and Sinhala BERT-based approach for handling transliteration ambiguity. Their system was explicitly designed to resolve cases where a single Romanized word may correspond to multiple distinct Sinhala meanings, using sentence-level context to make disambiguation decisions. Despite the progress made by these systems, they were evaluated primarily on general transliteration accuracy rather than on a benchmark explicitly targeting semantic disambiguation.

For Romanized Sinhala, Sumanathilaka et al. (2024) introduced the Swa-Bhasha dataset, which contains Romanized Sinhala–Sinhala transliteration pairs along with resources relevant to transliteration disambiguation. However, this dataset does not specifically address transliteration ambiguity in sentence-level contexts, nor does it offer a benchmark explicitly designed to evaluate context-sensitive disambiguation performance. While widely used transliteration datasets such as the Dhakshina Dataset (Roark et al., 2020) and Aksharantar (Madhani et al., 2022) exist for Indo-Aryan languages, they are limited in their support for transliteration disambiguation tasks.

Transliteration ambiguity is not unique to Sinhala; similar challenges are prevalent in other languages. In Hindi, for example, the Romanized word "sir" can correspond to either "सिर" (head) or "सर" (sir) depending on context, leading to semantic confusion in NLP applications (Kumar et al., 2025). Such issues are well-documented across Indo-Aryan languages, where word-level ambiguity in transliteration poses a significant challenge for machine learning models (Perera and Sumanathilaka, 2025). Despite their prevalence, existing resources in these languages largely focus on phonetic accuracy rather than sentence-level disambiguation, leaving a gap this work aims to address.

The transliteration disambiguation Singlish dataset proposed in this paper fills this gap by introducing the first resource specifically designed to evaluate how well transliteration systems handle semantic ambiguity in Romanized Sinhala. It draws inspiration from prior TD efforts and recent adversarial testing work, while addressing the unique challenges of ad-hoc Romanization and low-resource language settings.

3 Dataset Construction

This section describes the development of the TD dataset for Romanized Sinhala, including the motivation behind its creation, the methodology used to select ambiguous words, the sentence generation process, and the final dataset structure.

3.1 Motivation and Objectives

The primary goal of the dataset is to facilitate the evaluation of backward transliteration systems on their ability to resolve transliteration ambiguities in Romanized Sinhala, particularly in real-world, context-dependent usage. Romanized Sinhala words often lack one-to-one mappings with their native Sinhala counterparts, and users fre-

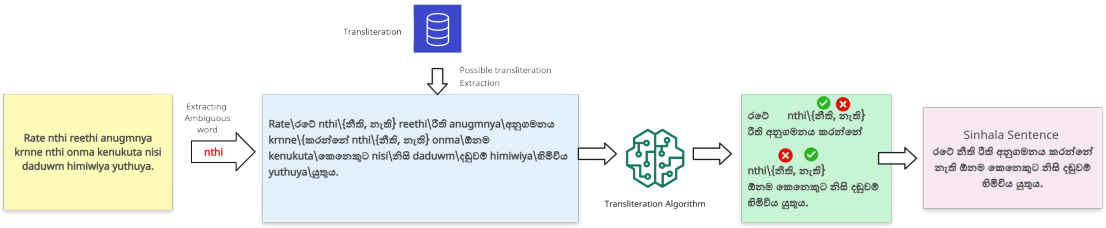

Figure 1: Transliteration Disambiguation process for the word *nthi*

quently write them in inconsistent, ad-hoc forms (Herath and Sumanathilaka, 2024). This results in situations where the same Romanized word can represent multiple Sinhala words with different meanings.

To address this, the dataset focuses on binary transliteration ambiguity—cases where a single Romanized form corresponds to two distinct Sinhala words. The dataset provides diverse contextual examples that force systems to rely on sentence-level semantics rather than surface forms alone.

3.2 Ambiguous Word Selection

The selection of ambiguous Romanized Sinhala words was carried out using an existing Romanized Sinhala–Sinhala dictionary compiled by Sumanathilaka et al. (2024). This dictionary provides mappings from ad-hoc Romanized Sinhala forms to their corresponding Sinhala script equivalents, reflecting the informal and varied nature of Romanization commonly seen in real-world usage.

To construct a high-quality TD dataset, the following filtering steps were applied to identify ambiguous Romanized words that map to exactly two distinct Sinhala meanings:

1. Spelling Verification: Sinhala entries in the dictionary were first checked for spelling correctness using the Madura Dictionary API (Madura-API, 2020). Entries with misspelled or invalid Sinhala words were removed to ensure linguistic accuracy.

2. Semantic Filtering: Some Romanized Sinhala words map to multiple Sinhala forms with similar or synonymous meanings (e.g., "ape" maps to "අපි", "අපේ", and "අප"). Such cases were excluded to ensure that each ambiguous word in the dataset corresponds to semantically distinct Sinhala meanings. Filtering was done using a verified lemma list (Fernando and Dias, 2021) and manual inspection for semantic contrast.

3. Binary Ambiguity Focus: Only Romanized Sinhala words with exactly two distinct, non-overlapping meanings were selected. Words with more than two meanings were discarded.

4. Frequency-Based Selection: To prioritize relevance and practicality, a Sinhala word frequency list (Fernando and Dias, 2021) was used. From the dictionary, Romanized Sinhala words were retained only if both of their corresponding Sinhala meanings appeared in the top 3,000 most frequent Sinhala words. This ensured that the selected words represent terms commonly encountered in everyday Sinhala usage.

After applying these filtering steps, a final list of 22 ambiguous Romanized Sinhala words was selected, as shown in Table 1. These words serve as the foundation for sentence generation in the dataset, each forming a minimal TD unit used to evaluate transliteration systems in ambiguous contexts.

3.3 Sentence Generation Process

For each ambiguous Romanized Sinhala word, 30 unique sentences were created to reflect different semantic contexts:

- 10 sentences each for one Sinhala ambiguous word
- 10 sentences containing both Sinhala ambiguous words within the same sentence

This three-fold structure supports evaluation of both isolated and compound contextual disambiguation in backward transliteration systems.

The sentence generation process was carried out in two main stages:

1. Sinhala Sentence Collection and Dual-Sense Generation: The process began by collecting

Singlish Word	Sinhala Meanings
nthi	නීති, නැති
yam	යම්, යාම
wsi	වාසි, වැසි
ud	උඩ, උදේ
thora	තොර, තෝරා
sthana	සටහන, ස්ථාන
samaya	සමය, සාමය
smawa	සමාව, සීමාව
ras	රස, රැස්
phra	පහර, පොහොර
oya	ඔය, ඔයා
nyk	නායක, නොයෙක්
mta	මට, මීට
maru	මරු, මාරු
mahath	මහත්, මහතා
es	ඇස, එසේ
eda	ඇද, එදා
dnna	දන්නා, දෙන්නා
deka	දැක, දෙක
bala	බල, බාල
badu	බඩු, බදු
adi	අඩි, ආදි

Table 1: Ambiguous Singlish words and their Sinhala meanings

Sinhala sentences from a publicly available dataset (Hettiarachchi et al., 2024). These sentences were selected such that each one clearly reflected one of the two Sinhala meanings associated with a target Romanized word. These collected sentences served as high-quality, real-world examples of single-sense usage.

To generate sentences that included both Sinhala meanings in a coherent context, the collected single-sense sentences were then passed as prompts to the Claude 3.0 Sonnet language model. This model was selected due to its strong support for the Sinhala language, making it well-suited for generating contextually appropriate and semantically accurate sentences (Jayakody and Dias, 2024). By presenting examples of each meaning, the model was guided to understand the distinct semantics of both Sinhala words and construct new sentences where both meanings naturally co-occur. This ensured that dual-sense sentences were not only grammatically valid but also semantically rich. The following prompt was used to generate dual-sense Sinhala sentences, each containing both ambiguous meanings in context. The generation was performed using a temperature of 0.2 and a maximum output length of 5000 tokens.

"A back-transliteration system has been developed to convert Romanized Sinhala text into native Sinhala script. This system is designed to accurately handle Romanized Sinhala words that correspond to multiple Sinhala meanings by leveraging contextual information. To evaluate this transliteration system, a test dataset must be created. The goal is to generate Sinhala sentences for specific ambiguous words. The ambiguous Romanized Sinhala word in focus is [singlish-word], which can correspond to either [sinhala-word1] or [sinhala-word2] in Sinhala.

Below are example sentences that include only one of the ambiguous meanings: Sentences containing [sinhala-word1]: [sentences1]

Sentences containing [sinhala-word2]: [sentences2]

Based on the examples above, generate 10 Sinhala sentences that include both [sinhala-word1] and [sinhala-word2] within the same sentence. Ensure that each sentence provides sufficient contextual information to help evaluate how well the transliteration system resolves ambiguity based on context."

2. Romanization Process: Once the full set of Sinhala sentences was finalized, they were converted into Romanized Sinhala using a rule-based transliteration system. For consistency, each target ambiguous word was manually replaced with its corresponding adhoc Romanized form from the Romanized Sinhala–Sinhala dictionary (Sumanathilaka et al., 2024). The remainder of the sentence was transliterated using consistent phonetic rules to preserve meaning and structure.

3. Manual Verification: Each created record was manually evaluated with the help of three linguists who are experts in the language and transliteration techniques. This process ensured that the dataset entries were accurately

Sentence Type	Example Sentence (Singlish / Sinhala)
Single Sense (වාසි)	**Singlish:** yamkisi pirisakata meya wsi sahagatha wana bawata owuhu chodana kara sitiyaha. **Sinhala:** යම්කිසි පිරිසකට මෙය වාසි සහගත වන බවට ඔවුහු චෝදනා කර සිටියහ.
Single Sense (වැසි)	**Singlish:** wsi jalaya ho apa jalaya gala yana pahath thenaka linda nokapanna. **Sinhala:** වැසි ජලය හෝ අප ජලය ගලා යන පහත් තැනක ළිඳ නොකපන්න.
Dual Sense	**Singlish:** wsi kalayedee gowithen katayuthuwalin uparima wsi laba geneemata naweena thakshanaya yoda gatha hekiya. **Sinhala:** වැසි කාලයේදී ගොවිතැන් කටයුතුවලින් උපරිම වාසි ලබා ගැනීමට නවීන තාක්ෂණය යොදා ගත හැකිය.

Table 2: Example sentences for the ambiguous word *wsi* from the TD dataset

selected and transliterated. Any incorrect sentences were revisited and corrected by the principal author of this work.

This combination of real-world scraping, controlled generation, and systematic Romanization ensures that the dataset is both linguistically diverse and realistically representative of how transliteration ambiguity appears in Romanized Sinhala.

3.4 Dataset Statistics and Format

The final TD dataset comprises a total of 660 sentence pairs, each consisting of a Romanized Sinhala sentence and its corresponding version in native Sinhala script. The dataset covers 22 ambiguous Romanized Sinhala words, each associated with two semantically distinct Sinhala meanings.

To facilitate targeted evaluation of backward transliteration systems, the dataset is organized into two distinct test sets:

- Test Set 1 – Single-Sense Sentences: This set contains Romanized Sinhala sentences where the ambiguous word is used in a context corresponding to only one of its possible Sinhala meanings. It is intended to evaluate a system's ability to infer the correct meaning based on sentence-level context.

- Test Set 2 – Dual-Sense Sentences: This set contains sentences where both Sinhala meanings of the ambiguous word appear within the same sentence. It is designed to assess the system's capacity to disambiguate and correctly transliterate each instance of the word in a more complex, semantically dense context.

Table 2 shows example sentences for the word 'wsi' from the TD dataset.

4 Evaluation Setup

To demonstrate the applicability of the proposed TD dataset, we conducted a comparative evaluation using four existing backward transliteration systems. The evaluation focuses on how effectively each system handles transliteration ambiguity in Romanized Sinhala backward transliteration process.

4.1 Systems Evaluated

To assess the effectiveness of the proposed TD dataset, we evaluated four existing backward transliteration systems for Romanized Sinhala. These systems vary in design, ranging from early rule-based approaches to more recent neural and hybrid models, offering a comprehensive performance comparison across different transliteration paradigms. Below are the systems used in the evaluation:

- Perera et al. (2025): This system adopts a hybrid approach that integrates an ad-hoc transliteration dictionary, rule-based mechanisms, and a BERT-based neural model specifically designed to handle transliteration ambiguities in Romanized Sinhala. The system was evaluated using two configurations: a base Sinhala BERT model [1] and a fine-tuned variant trained on a Sinhala text corpus [2]. This architecture is optimized for context-aware disambiguation during backward transliteration.

[1] https://huggingface.co/Ransaka/sinhala-bert-medium-v2/tree/main

[2] https://huggingface.co/Sameera827/Sinhala-BERT-MLM/tree/main

System	Model	Set	F1 Score
Perera et al. (2025)	Sinhala BERT	1	0.9601
		2	0.9209
	Fine-tuned BERT	1	0.9626
		2	0.9389
Dharmasiri and Sumanathilaka (2024)	GRU + Rule-based	1	0.3422
		2	0.3312
Sumanathilaka et al. (2023)	N-gram + Rule-based	1	0.3603
		2	0.3773
Real-Time Unicode Converter (2006)	Rule-based	1	0.3430
		2	0.3333

Table 3: Evaluation results of back-transliterators on the TD Dataset

- Dharmasiri and Sumanathilaka (2024): This approach introduces a hybrid model combining neural machine translation (NMT) and rule-based processing. A Gated Recurrent Unit (GRU)-based model forms the core of the neural component.

- Sumanathilaka et al. (2023): The proposed system uses a hybrid approach to transliterate Romanized Sinhala into native Sinhala script. It combines a trigram-based statistical model, a rule-based transliterator using defined phonetic patterns, and a Trie-based suggestion mechanism. The process begins with statistical prediction, followed by rule-based correction for ambiguous cases, and ends with word suggestions from a trained Trie structure.

- Real-Time Unicode Converter (2006): An early and widely-used rule-based transliteration system, this tool performs direct conversion from Romanized Sinhala to Sinhala Unicode script. It operates without contextual awareness or machine learning, relying solely on static character mappings, and serves as a baseline in this evaluation.

4.2 Evaluation Procedure

Each of the four back-transliteration systems was evaluated on the full set of Romanized Sinhala sentences from the TD dataset. The goal of the evaluation is to determine whether each system can accurately resolve the intended Sinhala meaning of an ambiguous Romanized word based on its context.

For each sentence, the system's output was compared against the reference Sinhala sentence provided in the dataset. A prediction is considered correct if the transliterated form of the ambiguous word matches the expected Sinhala meaning in the reference.

In the case of dual-sense sentences, where both Sinhala meanings are present in the same sentence, the system is expected to produce both correct forms within their appropriate context. Predictions involving Sinhala words outside the two predefined meanings for a given ambiguous word were treated as errors and excluded from F1-score calculation.

4.3 Evaluation Metrics

The evaluation uses the F1 score as the primary metric, balancing both precision and recall to measure transliteration accuracy. This metric reflects how effectively each system identifies the correct Sinhala meaning of an ambiguous Romanized word within a given context. F1 scores are reported separately for the two test sets:

- Test Set 1 measures the system's ability to perform accurate sense selection in straightforward contexts.

- Test Set 2 assesses the system's capacity to handle complex disambiguation involving dual interpretations of the same Romanized form.

By reporting performance across both single-sense and dual-sense scenarios, the evaluation highlights the strengths and limitations of each system in resolving transliteration ambiguity through backward transliteration.

System	Model	Output Sinhala Words
Perera et al. (2025)	Sinhala BERT	අඩි, අඩි
	Fine-tuned BERT	ආදි, අඩි
Dharmasiri and Sumanathilaka (2024)	GRU + Rule-based	අඩි, අඩි
Sumanathilaka et al. (2023)	N-gram + Rule-based	ආදි, ආදී
Real-Time Unicode Converter (2006)	Rule-based	අදි, අදි

Table 4: Model predictions for the ambiguous word `adi` in a dual-sense sentence. The expected Sinhala outputs are ආදි, අඩි.

5 Results and Analysis

This section presents the results of evaluating four back-transliteration systems on the TD dataset using F1 scores. Each system was tested on both Test Set 1 (sentences with a single intended Sinhala meaning) and Test Set 2 (sentences containing both meanings of the ambiguous word). The results are summarized in Table 3.

5.1 Observations

The best overall perf ormance was achieved by the fine-tuned BERT model from Perera et al. (2025), which obtained F1 scores of 0.9626 on Test Set 1 and 0.9389 on Test Set 2. The base Sinhala BERT also performed well, though slightly behind the fine-tuned model, showing that domain-specific tuning improves generalization.

Systems by Dharmasiri and Sumanathilaka (2024) and Sumanathilaka et al. (2023) showed significantly lower performance, with F1 scores ranging from approximately 0.33 to 0.37. While both incorporate hybrid approaches, the reliance on shallow statistical or sequence-based models may limit their effectiveness in resolving fine-grained contextual distinctions.

Real-Time Unicode Converter (2006), a purely rule-based system, performed comparably to the other traditional models, confirming the limitations of non-contextual transliteration in ambiguous settings.

5.2 Analysis

The results affirm the importance of context-aware neural approaches in back-transliteration tasks that involve semantic ambiguity. Systems using deep language models, particularly those trained or fine-tuned on relevant linguistic data, demonstrate a superior ability to resolve word sense based on sentence context. In contrast, systems that lack contextual modelling, whether rule-based or statistical, struggle to differentiate meanings, especially in Test Set 2, where both senses appear in a single sentence.

To better understand how back-transliteration systems handle context-sensitive ambiguity, we analyzed a dual-sense sentence from the TD dataset that includes the ambiguous Romanized Sinhala word “adi”. Table 4 compares each system’s transliteration output for both instances of the word, alongside the expected Sinhala forms. The correct outputs are ආදි (ancient) and අඩි (feet), in that order, as indicated by the sentence’s semantic context. The sentence used for this analysis is as follows: “adi kalaye sita ... gewathu wata adi hathak pamana usa ...”

These findings validate the usefulness of the proposed TD dataset as a benchmark for evaluating transliteration models’ ability to disambiguate based on context. They also underscore the gap between modern neural methods and earlier rule-based or statistical systems in this specific task.

5.3 General Transliteration Evaluation

While the core objective of this study is to evaluate transliteration disambiguation in ambiguous, context-sensitive settings using the proposed TD dataset, we also present results from existing back-transliteration systems on the IndoNLP dataset (Sumanathilaka et al., 2025) for complementary analysis. IndoNLP is a general-purpose benchmark that evaluates transliteration quality using surface-level metrics such as Word Error Rate (WER), Character Error Rate (CER), and BLEU score.

These results are not directly comparable to our TD dataset evaluations, as the two datasets serve different purposes. TD focuses on semantic disam-

System	Model	Set	WER	CER	BLEU
Perera et al. (2025)	Sinhala BERT	1	0.0888	0.0203	0.9113
		2	0.0917	0.0216	0.9084
	Fine-tuned BERT	1	0.0867	0.0200	0.9133
		2	0.0903	0.0215	0.9099
De Mel et al. (2025)	Rule-based	1	0.6689	0.2119	0.0177
		2	0.6809	0.2202	0.0163
	DL-based	1	0.1983	0.0579	0.5268
		2	0.2413	0.0789	0.4384
Dharmasiri and Sumanathilaka (2024)	GRU + Rule-based	1	0.3323	0.0827	0.6677
		2	0.4808	0.1567	0.5193
Sumanathilaka et al. (2023)	N-gram + Rule-based	1	0.2342	0.0542	0.7667
		2	0.2509	0.0678	0.7502

Table 5: Evaluation results of back-transliterators on the IndoNLP dataset

biguation of Romanized Sinhala words with multiple meanings, which IndoNLP does not capture. Instead, IndoNLP provides a useful baseline for assessing phonetic fidelity and overall transliteration correctness, independent of semantic ambiguity.

Table 5 reports the performance of several systems on IndoNLP. While systems such as the fine-tuned BERT model achieve high BLEU and low error rates, these results do not imply strong performance on ambiguity resolution. In fact, our TD evaluation highlights that systems performing well on IndoNLP may still struggle in disambiguating meaning in real-world sentences.

In conclusion, while traditional transliteration datasets remain useful for baseline assessment, they are not sufficient for fully evaluating the performance of back-transliteration systems. This reinforces the unique value of the TD dataset, which introduces ambiguity-focused challenges not addressed in existing Sinhala transliteration benchmarks.

6 Conclusion

This paper presented the first transliteration disambiguation dataset for Romanized Sinhala, aimed at evaluating backward transliteration systems in the presence of transliteration ambiguity. The dataset includes 22 ambiguous Romanized words, each with context-rich sentence pairs reflecting both Sinhala meanings. Evaluation across four transliteration systems showed that context-aware neural models significantly outperform rule-based and statistical methods in disambiguating meanings based on context. These results highlight the importance of contextual modeling for accurate back-transliteration. The dataset offers a valuable benchmark for Sinhala NLP and supports future work on improving transliteration. The TD dataset is publicly available at: https://github.com/Sameera2001Perera/Romanized-Sinhala-Transliteration-Disambiguation-Dataset

Limitations

The dataset is limited in its scope to binary lexical ambiguities, where each Romanized Sinhala word is associated with two semantically distinct Sinhala meanings. In reality, many Romanized Sinhala words—especially those written in ad-hoc, informal styles—can correspond to more than two Sinhala meanings or exhibit many-to-many mappings, where a single Sinhala word can be represented by multiple Romanized forms. To ensure semantic clarity, annotation feasibility, and high linguistic quality, we restricted the dataset to one-to-two mappings, using a filtering mechanism based on semantic distinction and frequency to retain only the most commonly used and semantically clear pairs. This design simplifies evaluation and focuses on clearly distinguishable ambiguities. Additionally, the dataset includes 22 ambiguous Romanized words, selected based on their frequency and relevance. While this provides a solid basis for evaluating context-sensitive disambiguation, the limited vocabulary size restricts the dataset's overall coverage. Expanding to include multi-sense ambiguity and many-to-many mappings remains a key direction for future work.

References

Maneesha U Athukorala and Deshan K Sumanathilaka. 2024. Swa bhasha: Message-based singlish to sinhala transliteration. *arXiv preprint arXiv:2404.13350*.

Yomal De Mel, Kasun Wickramasinghe, Nisansa de Silva, and Surangika Ranathunga. 2025. Sinhala transliteration: A comparative analysis between rule-based and Seq2Seq approaches. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 166–173, Abu Dhabi. Association for Computational Linguistics.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Sachithya Dharmasiri and T.G.D.K. Sumanathilaka. 2024. Swa bhasha 2.0: Addressing ambiguities in romanized sinhala to native sinhala transliteration using neural machine translation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 241–246.

Aloka Fernando and Gihan Dias. 2021. Building a linguistic resource : A word frequency list for Sinhala. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 606–610, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

H. Herath and Deshan Sumanathilaka. 2024. Tamzhi: Shorthand romanized tamil to tamil reverse transliteration using novel hybrid approach. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 17:1–7.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Ravindu Jayakody and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. In *2024 International Conference on Asian Language Processing (IALP)*, pages 162–167.

Saurabh Kumar, Dhruvkumar Babubhai Kakadiya, and Sanasam Ranbir Singh. 2025. Team IndiDataMiner at IndoNLP 2025: Hindi back transliteration - Roman to Devanagari using LLaMa. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 129–134, Abu Dhabi. Association for Computational Linguistics.

WMP Liwera and L Ranathunga. 2020. Combination of trigram and rule-based model for singlish to sinhala transliteration by focusing social media text. In *2020 From Innovation to Impact (FITI)*, volume 1, pages 1–5. IEEE.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Aksharantar: Open indic-language transliteration datasets and models for the next billion users. *arXiv preprint arXiv:2205.03018*.

Madura-API. 2020. madura-api.

Sandun Sameera Perera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka, and Isuri Anuradha. 2025. IndoNLP 2025 shared task: Romanized Sinhala to Sinhala reverse transliteration using BERT. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 135–140, Abu Dhabi. Association for Computational Linguistics.

Sandun Sameera Perera and Deshan Koshala Sumanathilaka. 2025. Machine translation and transliteration for indo-aryan languages: A systematic review. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 11–21.

Real-Time Unicode Converter. 2006. Real Time Unicode Converter.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing south asian languages written in the latin script: the dakshina dataset. *arXiv preprint arXiv:2007.01176*.

Deshan Sumanathilaka, Isuri Anuradha, Ruvan Weerasinghe, Nicholas Micallef, and Julian Hough. 2025. Indonlp 2025: Shared task on real-time reverse transliteration for romanized indo-aryan languages.

Deshan Sumanathilaka, Nicholas Micallef, and Ruvan Weerasinghe. 2024. Swa-bhasha dataset: Romanized sinhala to sinhala adhoc transliteration corpus. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 189–194.

T.G.D.K. Sumanathilaka, Ruvan Weerasinghe, and Y.H.P.P. Priyadarshana. 2023. Swa-bhasha: Romanized sinhala to sinhala reverse transliteration using a hybrid approach. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 136–141.