AsRED: Development and Evaluation of an Assamese Reduplication Dataset

Pankaj Choudhury[1], **Chaitanya Kirti**[1], **Dhrubajyoti Pathak**[1], **Sukumar Nandi**[1,2]
[1]Centre for Linguistic Science and Technology
[2]Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, Assam, India
{pankajchoudhury, ckirti, drbj153, sukumar}@iitg.ac.in

Abstract

Reduplication is a common linguistic phenomenon in many South Asian languages, including Assamese. The studies of reduplication are rich in literature in most languages. However, its study in low-resource languages with respect to computational linguistics is still lacking. In this paper, we introduce AsRED, a manually annotated dataset for reduplication detection in Assamese. The dataset covers diverse domains such as social media, news articles, textbooks, government websites, and wiki articles. The dataset consists of over 90K reduplicated tokens across 83K sentences. We evaluate the dataset using classical models like LSTM, BiLSTM, and CRF. We further incorporate contextual information from pretrained multilingual languages models like mBERT, XLM-R, MuRIL, and IndicBERT v2 to enhance performance. Experimental results demonstrate that the pre-trained multilingual language model MuRIL shows the best performance, achieving an F1-score of 0.9594. Furthermore, we present an error analysis of the best-performing model that highlights key challenges in reduplication detection. The error analysis further reveals specific linguistic properties of Assamese reduplication. The dataset and findings of this paper provide a foundation for further research on reduplication and morphological patterns in Assamese.

1 Introduction

Reduplication is a systematic repetition of any linguistic unit, such as a phoneme, morpheme, word, phrase, clause, or entire utterance. Reduplication (Hurch and Mattes, 2005) is a prevalent linguistic phenomenon observed in several languages (Rubino, 2013), including Indian languages. For Example, the word "*Bye-bye*" in English is a reduplicated version of the word "*Bye*". Other examples include "*Hip-hop*", "*Ping-pong*", "*Okey-dokey*" etc. It serves a unique grammatical and semantic purpose in a language. However, due to less attention in the context of text processing or automatic speech recognition, it is often mistaken for repetition or spelling error, leading to errors in the downstream Natural Language Processing (NLP) systems. Reduplicated word(s) carry a diverse range of semantic meanings and can occasionally serve as indicators of the speaker's emotional condition. It has many practical applications in various NLP tasks, such as sentiment analysis, Part-of-Speech (POS) tagging, Named Entity Recognition (NER), etc. Unfortunately, there has been little study into how these occurrences might be used to improve NLP tasks. It is one of the least studied NLP areas in languages with limited resources, such as Assamese. The primary factor contributing to this issue is the lack of availability of the dataset pertaining to reduplication. On the other hand, recent advances in Deep neural networks based systems (Vaswani et al., 2017; Rei et al., 2016), which achieve state-of-the-art results in various NLP tasks, require large datasets for a particular task. The existing work on Assamese reduplication is limited to linguistic studies only (Goswami, 2023). Considering the lack of resources for reduplication in Assamese language, the objective of this study is to provide an extensive dataset for Assamese reduplication.

Assamese language (Glottocode: assa1263) is an Indo-Aryan language and has similarities with other Indic languages such as Hindi, Bengali, Odia. Assamese has 15 million native speakers (Census, 2020 (accessed April, 2020) and is the official language of Assam, a state in North-east India. Assamese language makes extensive use of reduplication in comparison to other Indic languages (Goswami, 1987). As shown in Example 1, the word "লগে লগে" (/loge loge/) is an Assamese reduplicated word. Despite its relevance, reduplication in Assamese has received limited attention in NLP research.

1. চাৰিটা বজাৰ লগে লগে আমি যাম
   sarita bɔjar loge loge ami ʒam
   We will leave at four o'clock

To address the lack of resources for reduplication detection in Assamese, we compile a corpus from multiple domains including social media, school textbooks, government websites, and Wikipedia. In the absence of a well-structured Assamese corpus, we perform targeted web crawling to collect relevant textual data. The collected corpus is cleaned and manually annotated with reduplication labels at the token level. Our key contributions are as follows:

1. We introduce AsRED, a novel reduplication dataset for the low-resource Assamese language, comprising approximately 90K reduplicated tokens across 83K annotated sentences.
2. We provide a domain-wise analysis of reduplication patterns, covering sources such as newspapers, government websites, magazine articles, Wikipedia entries, social media posts, and educational materials.
3. We evaluate the dataset using standard sequence labeling models, including LSTM, BiLSTM, and CRF, and report results in terms of Precision, Recall, F1-score, and Accuracy.
4. We further enhance detection performance by incorporating contextual information from pretrained multilingual language models.

The rest of the paper is organized in the following manner. In section 2 a brief literature review has been given. section 3 describes various forms of reduplication present in the Assamese language. The proposed dataset development process is described in section 4. The section 5 is dedicated to the analysis of different statistics of the proposed dataset. The section 6 discusses the dataset evaluation, results and error analysis of the prediction model. Finally, the paper is concluded with a conclusion and future work in section 7.

2 Related Work

Reduplication is a widely researched phenomenon. Numerous typological research has been conducted in various forms of reduplication across many languages. On the other hand, very little work has been done in the field of text analytics to focus on recognition or model reduplication. Reduplication can be very useful to build many language tools. Beesley and Karttunen (Beesley and Karttunen, 2003; Roark and Sproat, 2007) used finite-state transducers (FST's) to compute reduplication. In their study, the authors modeled reduplication as a regular class of languages. However, some languages create copies of nouns as X-o-X while generating full reduplication, where X is a noun and -o- is an empty marker without semantic meaning. Hence, the reduplicated morpheme is unbounded. Dolatian et al. (Dolatian and Heinz, 2017) and Filiot et al. (Filiot and Reynier, 2016) introduced two-way finite-state transducers (2-way FSTs) for modelling reduplication. Later, Dolatian and Heinz (Dolatian and Heinz, 2019) created RedTyp a SQL database for different typological surveys of reduplication from 91 languages. Pathak et al. (2022) proposed a method for the identification of Assamese reduplication. They also present modeling of Assamese reduplication using 2-way FST. For other Indian languages like Bengali and Manipuri, automatic reduplication identification is covered as part of multiword expression (MWE)(Chakraborty and Bandyopadhyay, 2010; Nongmeikapam and Bandyopadhyay, 2011). Few works on Assamese reduplication are conducted by (Bora, 2016; Dattamajumdar, 2001), however, they are limited to solely descriptive linguistic studies. Considering this, we attempt to provide a large-scale reduplication detection dataset for the Assamese language. Moreover, we evaluated the dataset using deep learning based sequence classification models.

3 Assamese Reduplication

Reduplication in Assamese language can be classed into two types- (a) *Full Reduplication* and (b) *Partial Reduplication*.

3.1 Full Reduplication

In full reduplication, the entire word or word stem is repeated once (or twice in some cases) without any phonological change. In Example (2) the word "সিংহ" (/siŋhɔ/, 'Lion') is repeated two times as "সিংহই সিংহই" (/siŋhɔi siŋhɔi/) to put more emphasis that two or more "Lions" are fighting each other.

2. সিংহই সিংহই কাজিয়া কৰিছে
   siŋhɔi siŋhɔi kazia kɔrise
   The lions are fighting with each other

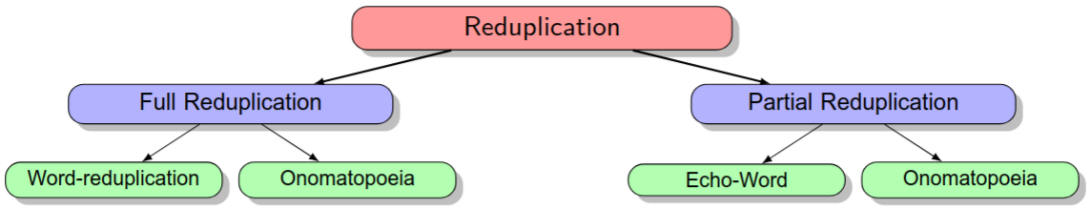

Figure 1: Different types of Assamese reduplication (Dattamajumdar, 1999)

As shown in Figure 1, the full reduplication can be further divided into *Word-reduplication* and *Onomatopoeic*. In Word-reduplication, words in different classes are reduplicated and change their semantic interpretation, as demonstrated in example (3). However, Onomatopoeic full reduplication represents sounds or senses that are used to express the language. Examples (4) and (5) show an Onomatopoeic full reduplication of the word "টং টং" (/tɔŋ tɔŋ/) and "গুণ গুণ" (/gun gun/). However, any single segment of Onomatopoeic full reduplication does not express semantic meaning (Bora, 2016).

3. তেওঁ মাজে মাজে খোজ কাঢ়িব যায়
   teo maze maze khoz karhibɔ zaj
   He sometimes goes for walk

4. ঘড়ীটো টং টং বাজিছে
   gʰorito tɔŋ tɔŋ bazise
   The clock is ringing

5. মৌ মাখীয়ে গুণ গুণ কৰে
   mou makʰiye gun gun kɔɹe
   Bee's are humming

3.2 Partial Reduplication

In partial reduplication the words have some phonological change in the second segment. Such reduplicated word is generated by simply copying the base element of the first segment prefix, suffix, or infix attached to it. As illustrated in Example (6) the word "ঘৰ" (/gʰɔr/, 'House') has a reduplicated segment "চৰ" (/sɔr/) by simply replacing "gʰɔ" with "sɔ".

6. বতাহে ঘৰ-চৰ ভাঙি পেলালে
   bɔtahe gʰɔr-sɔr bʰaŋi pelale
   The wind destroyed the house

The Partial Reduplication is further divided into *Echo-Word* and *Onomatopoeic*. In Echo-Word reduplication, the repeated part is created by a phonological copy of the base element with the addition of a suffix or prefix, and some partial alteration (see example 7). Here, the base element "মাছ" (/mas/, 'Fish') is a free morpheme and a lexical item of the language and "তাছ" (/tas/) is a prefix partial alteration of the base element. Alternatively, in Onomatopoeic partial reduplication, the second segment is created by copying and affixation, followed by the partial phonological alteration at any position of the repeated element. In example (8), the word "উখল মাখল" (/ukʰɔl makʰɔl/, 'Excitement') is the reduplicated word. The Onomatopoeic partial reduplication also represents feelings or senses and imitates natural sounds.

7. বজাৰত মাছ তাছ নাই
   bɔzarat mas tas naɪ
   There is no fish in the market

8. বিয়াৰ ৰভাত উখল মাখল লাগিছে
   biyar rɔbʰat ukʰɔl makʰɔl lagise
   There is a lot of excitement in the wedding reception

3.3 Semantic Features

There are several semantic features covered by reduplicated words. They are (a) Distributive Plurality, (b) Exclusiveness, (c) Degree of Manifestation (d) Similitude and (e) Reciprocity or Relationship.

(a) Distributive Plurality – The distributive plurality conveys plural form on an object. In the following example, the word "বছৰ বছৰ" (/bɔsɔɹ bɔsɔɹ/, 'For years') signifies that a certain "practice" ("প্ৰথা", /prɔtʰa/) has been going on for many years, which in turn is a plural form.

এইটো প্ৰথা বছৰ বছৰ ধৰি চলি আহিছে
eito prɔtʰa bɔsɔɹ bɔsɔɹ dʰɔri soli ahise
This practice has been going on for years

(b) Exclusiveness – Exclusiveness is used to mean something associated with only a select group or person. In the following example, the

reduplicated word "ধনী ধনী" (/dʰɔni dʰɔni/, 'Rich people') emphasizes the exclusiveness.

এই ঠাইলৈ ধনী ধনী মানুহ আহে
ei tʰailɔɪ dʰɔni dʰɔni manuh ahe
Rich people come to this place

(c) Degree of Manifestation – The Degree of Manifestation is used to convey different degrees of 'incompletion', 'excellence', 'mildness', 'intensity', 'hesitation' etc. In the following example, the word "লাহে লাহে" (/lahe lahe/, 'Very slowly') means the intensity of the speed of cars.

গাড়ীবোৰ বহুত লাহে লাহে গৈ আছে
garibore bahut lahe lahe gɔi ase
The cars are moving very slowly

(d) Similitude – Some reduplication is often used to show similarity or comparison. In the following example, the word "ৰজা ৰজা" (/ɹɔza ɹɔja/) means that the person compares himself with a "king".

তেওঁ আজিকালি ৰজা ৰজা যেন ভাবত থাকে
teo azikali ɹɔza ɹɔja bʰavat tʰake
He acts like a king these days

(e) Reciprocity or Relationship – Reciprocity semantic features in reduplication are to express the mutual relationship. In the following example, the word "গাড়ীয়ে গাড়ীয়ে" (/garije garije/) represents a relationship between "Cars".

গাড়ীয়ে গাড়ীয়ে খুন্দা লাগিছে
garije garije kʰunda lagise
Cars are colliding with each other

4 Dataset Development

The source corpus and preparation of the reduplication dataset are described in section in Section 4.1 and 4.2.

4.1 Corpus collection

There are only a few curated, monolingual corpora available online for the majority of Indian languages. The corpus of the Assamese language is less in size when compared to other languages spoken in India. One of the fundamental challenges in doing text analysis in Assamese is the gathering of a text corpus. In Assamese, reduplication is widely used in everyday conversation, poetry, and literature. It would be a fair study in reduplication distribution if it is done on text from different domains. Hence, we crawled a corpus of Assamese text from various domains in order to analyze the occurrence of reduplication in these domains and extract the reduplicated words. The domains comprise articles from Newspapers, Magazines, the Public Information Bureau (PIB), the Government websites, the Prime Minister's Speech, Wikipedia, Social media, High school textbooks, Health, and Culture. The statistics of the collected corpus is listed in Table 1. The Assamese newspaper, Niyomiya Barta corpus, is the largest, followed by PIB corpus, Wikipedia, and others. The corpora have a total of approximately 1.75 million sentences and 21.5 million tokens. A quality check is performed on the sentences comprising the corpora to ensure the absence of duplicate sentences.

4.2 Reduplication dataset development

Identifying reduplication occurrences in sentences was a key step in the dataset preparation process. We employed a rule-based reduplication identification system introduced in (Pathak et al., 2022). This system, originally developed using 0.11 million sentences from three domains (which are not part of the current corpus). We used the system as a recommender to assist us during annotation by suggesting possible reduplicated words. This helped streamline the annotation process. The annotations were performed by the co-authors of this paper, who are native speakers of Assamese. One co-author who is bilingual acted as an independent evaluator. To evaluate inter-annotator agreement (IAA), we randomly selected 1,000 sentences and each annotator independently labeled reduplications. This process resulted in an IAA score of 94.8%. Following this evaluation, the full corpus was annotated accordingly.

5 Dataset Statistics and Analysis

This section provides an analysis of the reduplication dataset from various perspectives, along with a summary of statistics pertaining to the occurrence of reduplication. Table 2 reports the statistics of the reduplication occurrence for each domain individually. The number of sentences with reduplication in each domain is presented, as well as the percentage of reduplication occurrence out of all sentences in that domain. The findings reveal that the domain of literature pertaining to Assamese Culture has the highest occurrence of reduplication, accounting for 8.72% of the total sentences. It suggests that reduplication is more prevalent in articles

Table 1: Details of corpora of various domains

Corpus	Category	Total Tokens	Total sentences
PM Speech	Govt speech	443K	31K
PIB India	Press Information Bureau	2004K	131K
Niyomiya barta	News Paper	4310K	279K
Asomiya Pratidin	News Paper	6481K	484K
School textbook	Textbook	194K	29K
Monikut	Magazine	894K	88K
Vikaspedia (Health)	Article	2409K	199K
Vikaspedia (Assamese Culture)	Article	995K	82K
Wikipedia	Wiki	3427K	385K
Social Media	Misc	325K	39K
Total		21482K	1747K

Table 2: Statistics of reduplication dataset

Corpus	Total Sentence	Sentence with reduplication	Reduplicated word	% total sentences
PM Speech	31K	2043	2291	7.08%
PIB India	131K	7090	7618	5.81%
Niyomiya barta	279K	13191	13969	5.00%
Asomiya Pratidin	484K	22106	23671	4.9%
School textbook	29K	644	687	2.37%
Monikut	88K	6341	6940	7.87%
Vikaspedia (Health)	199K	12556	13701	6.88%
Vikaspedia (Assamese Culture)	82K	6306	7173	**8.72%**
Wikipedia	385K	10355	11003	2.87%
Social Media	39K	2723	2978	7.58%
Total	1747K	83K	90K	5.15%

related to Assamese cultures.

In contrast, we noted that the occurrence of reduplication is the least prevalent in the School textbook and Assamese Wikipedia articles, accounting for 2.37% and 2.87% of the total sentences, respectively. The corpus of the class textbook comprises mathematical terminology, formulae, and scientific articles that employ fewer reduplicated words. Hence, it is evident that the statistical analysis results are valid. The Wikipedia text comprises articles from different areas, and the reduplication occurrence is less in those texts.

The rate of reduplication occurrence in newspaper articles from the two prominent Assamese newspapers, Niyomiya Barta and Asomiya Pratidin, is nearly identical at 5% and 4.9%, respectively. The PIB corpus is similar to newspapers, which serve as the Government of India's primary agency responsible for transmitting information to both print and electronic media platforms about government policies, programs, initiatives, and accomplishments. It consists of 5.81% reduplicated words throughout its articles. The magazine, health, and social media articles or stories have almost a similar rate of reduplication occurrence.

Table 3 presents the most frequently occurring reduplicated words across various domains. These reduplications reflect common linguistic patterns and semantic groupings prevalent in the language. For instance, the word "কাম-কাজ" (/kam-kaz/, 'work and others') appears prominently in multiple corpora such as PM Speech, Niyomiya Barta, Wikipedia, and Vikaspedia (Health), indicating its widespread usage in formal and informational contexts. Similarly, words like "মুখামুখি" (/mukhamukhi/, 'face to face'), "ৰীতি-নীতি" (/riti-niti/, 'customs and traditions'), and "লৰালৰি" (/lɔralɔri/, 'to move or act quickly') showcase the diversity of semantic domains ranging from social interaction to cultural expression and physical movement captured through reduplication. The recurrence of certain patterns across domains also highlights the functional and stylistic roles reduplication plays in Assamese discourse.

6 Dataset Evaluation

6.1 Task Description

Reduplication detection involves identifying repeated or partially repeated word patterns within a sentence. We formulate this task as a token-level sequence labeling problem. The labeling follows the BIO tagging scheme, where tokens are tagged

Table 3: Top Reduplicated Words Across Corpora

Corpus Name	Top Reduplicated Word
PM Speech	কাম-কাজ (/kam-kaz/, 'work and others')
PIB India	বুজাবুজি (/buzabuzi/, 'mutual understanding')
Niyomiya Barta	কাম-কাজ (/kam-kaz/, 'work and others')
Asomiya Pratidin	মুখামুখি (/mukhamukhi/, 'Face to face')
School Textbook	নদ-নদী (/nɔd-nɔdi/, 'Rivers and rivulet')
Monikut	কিবাকিবি (/kibakibi/, 'something something')
Vikaspedia (Health)	কাম-কাজ (/kam-kaz/, 'work and others')
Vikaspedia (Culture)	ৰীতি-নীতি (/ɹiti-niti/, 'customs and traditions')
Wikipedia	কাম-কাজ (/kam-kaz/, 'work and others')
Social Media	লৰালৰি (/lɔralɔri/, 'to move or act quickly')

as beginning (**B**), inside (**I**), or outside (**O**) of a reduplicative expression. Given an input sentence $\mathbf{S} = [w_1, w_2, \ldots, w_n]$ of n tokens, the model outputs a label sequence $\mathbf{L} = [l_1, l_2, \ldots, l_n]$, with each $l_i \in O, B\text{-}RED, I\text{-}RED$. Here, *B-RED* marks the start of a reduplicative unit, *I-RED* marks its continuation, and *O* denotes tokens not part of any reduplication.

6.2 Evaluation Models

We evaluate the proposed reduplication detection dataset using classical sequence labeling models, including Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Conditional Random Field (CRF), and BiLSTM with CRF (BiLSTM+CRF). Among these, BiLSTM achieved the best performance and was selected as the baseline encoder. Additionally, we have used several encoder-only transformer models to capture contextual information to further improve performance. These models are pretrained on the Masked Language Modeling (MLM) objective. The encoder-only transformer models are multilingual BERT (mBERT) (Kenton and Toutanova, 2019), XLM-RoBERTa (XLM-R) (Conneau et al., 2020), Multilingual Representations for Indian Languages (MuRIL) (Khanuja et al., 2021), and IndicBERT v2 (Doddapaneni et al., 2023). mBERT is trained on Wikipedia data from 104 languages, enabling cross-lingual generalization. XLM-R is trained on 2.5TB of CommonCrawl data from 100 languages. It performs better than mBERT, especially for low-resource languages. MuRIL is designed for Indian languages, trained on both monolingual and parallel corpora in 17 languages. IndicBERT v2, built on the ALBERT architecture (Lan et al., 2019). It is trained on IndicCorp v2 (Doddapaneni et al., 2023) dataset, which is a monolingual corpus of 20.9 billion tokens and 1.1 billion sentences. IndicBERT v2 supports 24 Indian languages and performs well on the IndicXTREME benchmark (Doddapaneni et al., 2023).

6.3 Results and Discussion

Table 4: Performance of classical and transformer-based models.

Model	P	R	F1	Accuracy
LSTM	0.9149	0.7520	0.8255	0.7028
CRF	0.7820	0.4140	0.5413	0.3711
BiLSTM+CRF	0.9441	0.8610	0.9006	0.8192
BiLSTM	0.9365	0.8780	0.9063	0.8287
+ IndicBERT v2	0.9593	0.9550	0.9572	0.9178
+ mBERT	0.9189	0.7754	0.8411	0.7257
+ XLMR	0.9354	0.8952	0.9149	0.8431
+ MuRIL	**0.9612**	**0.9575**	**0.9594**	**0.9219**

Table 4 presents the performance of various models on the reduplication detection task. The performance of the task is reported in terms of Precision (P), Recall (R), F1-score (F1), and Accuracy. We evaluate both classical sequence labeling models and pretrained language model (PLM)-based approaches.

Among the classical models, the BiLSTM achieves the highest F1-score of 0.9063, outperforming LSTM (F1: 0.8255), CRF (F1: 0.5413), and BiLSTM+CRF (F1: 0.9006). The relatively poor performance of the CRF model can be attributed to its limited contextual representation, as it relies heavily on handcrafted features and lacks deep contextual encoding. While combining BiLSTM with CRF improves results compared to CRF alone, it still performs slightly below the standalone BiLSTM in terms of both F1 and accuracy. This suggests that for this task, deep contextual representations provided by BiLSTM alone are sufficient and the CRF layer does not add significant benefit.

To further enhance performance, we use the BiLSTM architecture as the base sequence labeler

and augment it with contextual embeddings from four PLMs. All PLMs show substantial improvements over the classical BiLSTM. The best performance is achieved using BiLSTM + MuRIL, with a precision of 0.9612, recall of 0.9575, F1-score of 0.9594, and accuracy of 0.9219. This result demonstrates the effectiveness of MuRIL in capturing reduplication patterns in Indian languages. This may be due to its pretraining on diverse Indian corpora including both monolingual and parallel data. IndicBERT v2 also performs competitively, achieving an F1-score of 0.9572, highlighting the strength of Indic-specific PLMs. XLM-R, while being trained on a large-scale multilingual corpus, also yields strong results (F1: 0.9149), showing good generalization. In contrast, mBERT, though still outperforming classical models, lags behind other PLMs (F1: 0.8411). This is possibly due to its limited capacity and training data compared to XLM-R and MuRIL.

Overall, the results clearly demonstrate that incorporating PLMs significantly boosts the performance of reduplication detection. Among these, MuRIL proves most effective, likely due to its focus on Indian languages and better handling of linguistic diversity and morphological complexity. These findings support the use of PLM-based encoders in sequence labeling tasks involving low-resource or morphologically rich languages.

6.4 Error Analysis

We conduct a detailed error analysis of the BiLSTM+MuRIL model to understand its limitations in reduplication detection. Based on the evaluation, we identify four major types of errors, illustrated through representative examples in Tables 5, 6, 7 and 8.

False positives from non-reduplicative wordforms. Some words in Assamese contain internal repetition or phonological patterns that resemble partial reduplication but are not semantically reduplicated. For example, words like "হাওঁফাওঁ" (/hɑõpʰɑõ/, 'lung') or "মুখামুখি" (/mukhamukhi/, 'face to face') are incorrectly labeled as reduplications by the model (Table 5). These forms exhibit surface-level similarity with reduplicated tokens but are in fact atomic lexical items, leading to over-prediction.

Errors from transliterated terms. The model also mistakenly tags transliterated English words as reduplicated due to character-level repetition or syllabic patterns. As shown in Table 6, terms such as "ডিআৰডিঅ" (/diɑrdio/, 'DRDO'), "য়ুনিয়ন" (/juniɔn/, 'Union'), and "চি চি" (/c c/, 'CC') are incorrectly identified as reduplications. This indicates the model's sensitivity to repetitive sequences, regardless of origin or linguistic function.

Table 5: Examples of false positives from non-reduplicative wordforms error.

Word	IPA	English
হাওঁফাওঁ	/hɑõpʰɑõ/	'lung'
মুখামুখি	/mukhamukhi/	Face to face or facing each other

Table 6: Examples of errors due to transliterated terms.

Word	IPA	English
ডিআৰডিঅ	/diArdio/	'DRDO' - Defence Research and Development Organisation
য়ুনিয়ন	/juniOn/	Union
চি চি	/c c/	CC - Cubic centimeter

Ambiguity in compound structures. Assamese often uses hyphenated compound words to convey inclusiveness or plurality, such as "ছাত্ৰ-ছাত্ৰী" (/satrɔ-satri/, 'male and female students') or "মাছ-মাংস" (/mas-maŋkʰɔ/, 'Fish and meat'). While these constructions exhibit semantic symmetry and surface repetition, they are syntactic compounds rather than true reduplications. As shown in Table 7, the model frequently labels such cases as reduplications, which raises ambiguity around the definition boundary between coordination and reduplication.

Table 7: Examples of errors due to ambiguity in compound structures.

Word	IPA	English
ছাত্ৰ-ছাত্ৰী	/satrɔ-satri/	Students (Male and Female)
যান-বাহন	/jan-vahan/	'All type of vehicles'
কাম-কাজ	/kam-kaz/	'Works and others'
মাছ-মাংস	/mas-maŋkʰɔ/	'Fish and meat'

Morphological challenges with inflected forms. While the model performs reasonably well on base reduplicated forms, it struggles with identifying reduplication when inflectional suffixes are attached. For example, in the word "বুজাবুজি" (/buzabuzi/, 'mutual understanding'), the reduplication is correctly detected. However, in its inflected forms such as "বুজাবুজিৰ" (/buzabuzir/, 'For mu-

tual understanding') or "বুজাবুজিত" (/buzabuzit/, 'In mutual understanding'), the model often fails to identify the base reduplication due to the added morphemes 'ৰ' (/r/) or 'ত' (/t/). This suggests a need for improved morphological robustness in the model's token representations.

Table 8: Examples of error due to morphological challenges with inflected forms

Word	IPA	English
বুজাবুজিৰ	/buzabuzir/	'For mutual understanding'
বুজাবুজিত	/buzabuzit/	'In mutual understanding'

7 Conclusion

In this work, we present AsRED, a manually annotated dataset for reduplication detection in Assamese. The dataset spans multiple domains, including social media, news, textbooks, and government websites. The dataset comprises over 90K reduplicated tokens across 83K sentences. We formulate reduplication detection as a sequence labeling task and evaluate the dataset using classical models such as LSTM, BiLSTM, and CRF. To incorporate contextual information and improve performance, we further experiment with encoder-based pretrained language models, including mBERT, XLM-R, MuRIL, and IndicBERT v2. The results show that pretrained multilingual models, especially MuRIL, achieve better performance compared to classical baselines. We also conduct an error analysis of the best-performing model for reduplication detection. We identify common challenges encountered by the models during reduplication detection. These include false positives from transliterated and compound words, and difficulties in recognizing reduplications in morphologically inflected forms. The analysis highlights specific linguistic properties of Assamese reduplication and points to future directions such as integrating morphological processing and syntactic features to improve detection accuracy.

Limitations

While this work introduces a novel dataset and provides a comprehensive evaluation of reduplication detection in Assamese, it has certain limitations. First, the annotation focuses primarily on surface-level reduplication patterns and may not capture deeper syntactic or semantic variations. Second, the current formulation treats reduplication as a flat sequence labeling task, which may overlook hierarchical or multi-word expressions. Additionally, the models show reduced performance when handling morphologically inflected reduplications or context-dependent reduplicative constructions. Finally, while we include data from multiple domains, the dataset may still not fully represent the diversity of Assamese usage across dialects, informal speech, or creative writing.

Ethical Considerations

This work involves the creation and annotation of a reduplication detection dataset for the Assamese language. All data used in this study were collected from publicly available sources, including government websites, online newspapers, educational materials, and social media platforms. We ensured that no personally identifiable information (PII) or sensitive content was included in the dataset. The annotations were performed manually by native speakers with linguistic expertise.

As Assamese is a low-resource language with cultural and linguistic diversity, we acknowledge the risk of underrepresenting certain dialects or usage patterns. We encourage future work to expand coverage to more dialects and speaker communities. The dataset is intended solely for research and educational purposes, and we caution against its use in applications that could lead to unintended social or cultural biases.

References

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. csli.

L Saikia Bora. 2016. Assamese grammar and usage: An analytical studies of assamese grammar and usage. *Pan-bazar, Guwahati: Chandra Prakash, Guwahati*.

Census. 2020 (accessed April, 2020). *ABSTRACT OF SPEAKERS' STRENGTH OF LANGUAGES AND MOTHER TONGUES - 2011*.

Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of reduplication in bengali corpus and their semantic analysis: A rule based approach. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 73–76.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Satarupa Dattamajumdar. 1999. *A contrastive study of the reduplicated structures in Asamiya Bangla and Odia*. Ph.D. thesis, Department of Linguistics, University of Calcutta, Kolkata, West Bengal.

Satarupa Dattamajumdar. 2001. A contrastive study of the reduplicated structures in asamiya, bangla and odia. *(No Title)*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Hossep Dolatian and Jeffrey Heinz. 2017. Reduplication with finite-state technology. *Proc. CLS*, 53:55–69.

Hossep Dolatian and Jeffrey Heinz. 2019. Redtyp: A database of reduplication with computational models. *Proceedings of the Society for Computation in Linguistics*, 2(1):8–18.

Emmanuel Filiot and Pierre-Alain Reynier. 2016. Transducers, logic and algebra for functions of finite words. *ACM SIGLOG News*, 3(3):4–19.

G. C Goswami. 1987. *Fundamentals of Assamese Grammar (* ɔsomiya byɑkʉrʉnʉr mɔʊlik bichɑr*, 11th Edition(Reprint,2017)*. Bina Library, Panbazar, Guwahati.

Vikas Goswami. 2023. Unlocking assamese derivational morphology: A comprehensive exploration of lexical word categories. *American Journal of Philological Sciences*, 3(09):05–11.

Bernhard Hurch and Veronika Mattes. 2005. *Studies on reduplication*. 28. Walter de Gruyter.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Kishorjit Nongmeikapam and Sivaji Bandyopadhyay. 2011. Identification of reduplicated mwes in manipuri: A rule based approach. In *Proceedings of 23rd International Conference on the Computer Processing of Oriental Languages (ICCPOL'10)*, pages 49–54.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. Reduplication in Assamese: Identification and Modeling. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Marek Rei, Gamal Crichton, and Sampo Pyysalo. 2016. Attending to Characters in Neural Sequence Labeling Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318, Osaka, Japan. The COLING 2016 Organizing Committee.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*, volume 4. OUP Oxford.

Carl Rubino. 2013. Reduplication. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.