Zero-shot Disfluency Detection for Indian Languages

Rohit Kundu, Preethi Jyothi and Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{rkundu, pjyothi, pb}@cse.iitb.ac.in

Abstract

Disfluencies that appear in the transcriptions from automatic speech recognition systems tend to impair the performance of downstream NLP tasks. Disfluency correction models can help alleviate this problem. However, the unavailability of labeled data in low-resource languages impairs progress. We propose using a pretrained multilingual model, finetuned only on English disfluencies, for zero-shot disfluency detection in Indian languages. We present a detailed pipeline to synthetically generate disfluent text and create evaluation datasets for four Indian languages: Bengali, Hindi, Malayalam, and Marathi. Even in the zero-shot setting, we obtain F1 scores of 75 and higher on five disfluency types across all four languages. We also show the utility of synthetically generated disfluencies by evaluating on real disfluent text in Bengali, Hindi, and Marathi. Finetuning the multilingual model on additional synthetic Hindi disfluent text nearly doubles the number of exact matches and yields a 20-point boost in F1 scores when evaluated on real Hindi disfluent text, compared to training with only English disfluent text.

1 Introduction

Disfluencies (e.g., filled pauses, repetitions, discourse markers) are artefacts that are inherent to spontaneous or conversational speech. Disfluencies typically obey the following surface structure comprising: a reparandum, an interruption point (+) that marks the end of the reparandum, an interregnum, and finally the repair (Shriberg, 1994). The reparandum consists of one or more words that are not intended by the speaker and will be replaced or ignored. The interregnum consists of an editing term indicating that the reparandum will be edited, or it may be empty, or it can contain fillers, discourse markers, etc. The repair section reflects the fluent part of the utterance. Words from the reparandum are repeated or corrected in the repair section, or a new chain of thought is started in case of a false start.

Consider the following example that illustrates two disfluency types:

{well} [i think + {you know} i think] the idea will work

The words highlighted in red and green refer to discourse marker and repetition disfluency types, respectively. The part in blue is the fluent version of the original sentence. The example also follows the standard annotation scheme:

[reparandum + {interregnum} repair]

Disfluencies in automatically transcribed text pose a major challenge for downstream NLP tasks such as machine translation, summarization, etc. (Rao et al., 2007; Wang et al., 2010). Disfluency detection/correction is often used as a preprocessing step for NLP, where the goal is to identify/remove the disfluent words (Shriberg et al., 1992). While disfluency correction has been extensively studied for English (Honal and Schultz, 2003; Zayats et al., 2014), it has received far less attention in other languages. This is largely due to the lack of labeled data for other languages.

In this work, our main objective is to build disfluency detection models for four Indian languages — Bengali, Hindi, Malayalam, and Marathi — in the zero-shot setting with no access to labeled disfluent data in these languages. To the best of our knowledge, this is the very first study of disfluency detection across multiple Indian languages and also the very first to investigate the ability of large pretrained models to do zero-shot disfluency detection.

We specify a rule-based procedure to generate disfluencies starting from fluent sentences. It is worth noting that the synthetically generated disfluent data might not completely reflect real world disfluencies. Nevertheless, we find the synthetic data to be useful in improving disfluency detection.
for low-resource Indian languages. Also, we can create near-real disfluent data by manually editing the synthetic data, which will take significantly less time than annotating from scratch. Our rule-based pipeline is targeted at Indian languages, and the same set of rules is applied to sentences in four Indian languages — Bengali, Hindi, Malayalam, and Marathi. From within these synthetic datasets, native speakers of the respective languages manually identified disfluent sentences that seemed natural. This resulted in manually-verified evaluation datasets for all the four languages. We also constructed evaluation datasets for Bengali, Hindi, and Marathi with real\(^1\) disfluent sentences, transcribed and extracted from conversational speech in interviews. Using these datasets as evaluation benchmarks and inspired by prior work on cross-lingual zero-shot transfer using large pretrained multilingual models (Pires et al., 2019; Hu et al., 2020; Khanuja et al., 2021), we investigated the effectiveness of a large pretrained multilingual model MuRIL (Khanuja et al., 2021) on the task of zero-shot disfluency detection. MuRIL is a multilingual transformer-based model that is pretrained on large amounts of text in a number of different Indian languages. We finetuned MuRIL using labeled disfluent sentences in English (Godfrey et al., 1992) (and synthetic disfluent text) and evaluated disfluency correction for all four Indian languages in the zero-shot setting.\(^2\)

Four MuRIL-based disfluency detection models were trained, viz. those using (1) only real English disfluent data, (2) both real English and synthetically generated Bengali disfluent data, (3) both real English and synthetically generated Hindi disfluent data, and (4) both real English and synthetically generated Marathi disfluent data. Model (2) significantly improves disfluency detection on the real Bengali disfluent data compared to model (1). We also observe similar results for the other two languages. This validates our claim that our synthetically generated data is effective in capturing (some subset of) the kinds of disfluencies that are encountered in real conversational data.

The main ideas in this work can be summarized as follows:

- We outline a common rule-based procedure that allows us to synthesize disfluencies for four Indian languages: Bengali, Hindi, Malayalam, and Marathi.

- We construct manually-verified evaluation datasets for all four Indian languages, starting from synthetically generated data. The Bengali, Hindi, Malayalam, and Marathi test sets contain 500, 575, 575, and 420 sentences, respectively.

- We also annotate real labeled disfluent datasets in Bengali, Hindi, and Marathi, containing 300, 150, and 250 sentences, respectively. These sentences were transcribed and extracted from real conversational speech. We note that this annotation process is substantially more tedious than identifying natural disfluencies starting from our synthetic data.

- We finetune a pretrained multilingual model, MuRIL, on labeled disfluent data in English and show its effectiveness at zero-shot disfluency detection for all four Indian language datasets.

- We show the utility of our synthetic disfluency generation pipeline by comparing performance of a model finetuned only on real English disfluent data versus a model finetuned on both real English and synthetically generated disfluent data of one Indian language.

- We present a detailed breakdown of performance across various disfluency types, show qualitative analyses of our model predictions and highlight some interesting aspects related to disfluencies in Indian languages (e.g., reduplication).

2 Related Work

There are three main categories of approaches for disfluency detection. They are based on (1) sequence tagging, (2) parsing, and (3) a noisy channel model (Kundu et al., 2022).

Sequence tagging based approaches use classification techniques to label individual words (Liu et al., 2006; Ostendorf and Hahn, 2013; Zayats et al., 2014; Ferguson et al., 2015; Hough and Schlangen, 2015; Zayats et al., 2016; Wang et al.,

\(^1\)We use the term real throughout the paper to contrast the manually-edited synthetic datasets with the (real) datasets containing disfluencies annotated from conversations.

\(^2\)Our code and datasets can be found at https://github.com/RKKUNDU/zero-shot-disfluency-detection.
We focus on four major disfluency types as listed in Appendix where various disfluencies are introduced in a fluent and apply to Indian languages. The goal is to find the most likely fluent sentence to introduce disfluencies in fluent sentences. We total of nine rules across the four disfluency types given has been added, resulting in a disfluent sentence Y. The main idea behind a noisy channel model of disfluency is that we assume there is a fluent source sentence X to which some noise has been added, resulting in a disfluent sentence Y. The goal is to find the most likely fluent sentence given Y (Johnson and Charniak, 2004; Zwarts and Johnson, 2011; Jamshid Lou and Johnson, 2017).

Prior works (Hu et al., 2020; Khandu et al., 2021) have used pretrained multilingual models for many zero-shot NLP tasks such as Named Entity Recognition (NER), Part of Speech (POS) tagging, Question Answering (QA), etc. However, this is the first work to attempt disfluency detection in a zero-shot setting and the very first work to study disfluency detection for multiple Indian languages.

Our work on synthetic disfluency data generation has parallels to the recent work of Passali et al. (2022) where they focus on an artificial disfluency generation algorithm. They focus broadly on Repetitions, Replacements, and Restarts and only focus on English. Saini et al. (2020) is another prior work that has looked into inducing disfluencies in English fluent text. Our disfluencies are much more fine-grained in construction (e.g., pronoun corrections, missing syllables, etc.) compared to prior work and apply to Indian languages.

## 3 Generating Synthetic Disfluencies

We focus on four major disfluency types as listed in Honal and Schultz (2003), i.e., Fillers, Repetitions, Corrections, and False Starts. We specify a total of nine rules across the four disfluency types to introduce disfluencies in fluent sentences. We show examples of Bengali disfluent sentences in Appendix A for all the disfluency types. Apart from what we describe in this section, there are some more fine-grained details governing how and where various disfluencies are introduced in a fluent sentence; these details are specified in our released codebase.

### 3.1 Fillers

We loosely use the term Fillers to denote editing terms, discourse markers, filled pauses and interjections. Editing terms are used to explicitly indicate that the previously uttered word(s) were not intended. Discourse markers help in beginning or keeping a turn (e.g., well) or merely serve as a form of acknowledgment (e.g., yeah). Filled pauses are non-lexicalized sounds without any semantic content. Interjections are defined as non-lexicalized sounds indicating affirmation or negation.

We simply introduce frequent filler phrases at randomly chosen positions. We choose frequent filler phrases after carefully observing conversations. We assume that there will be at most 3 fillers in a sentence and uniformly choose a number between 1 and 3. Thereafter, with uniform probability, we pick the location in the sentence at which the next filler will be inserted and also choose the filler phrase to be inserted with uniform probability from a pool of filler phrases.

Speakers might tend to use fillers before long words. For words with 12 or more characters, we first choose a filler phrase with uniform probability and then place it before the long word.

### 3.2 Repetitions

Repetition is defined as the phenomenon of speakers repeating a word or phrase.

**Word Repetition.** For this rule, we pick a word uniformly at random and repeat it.

**Phrase Repetition.** In this rule, we repeat a phrase containing 2 to 5 words. We first randomly pick a length from [2, 3, 4, 5] using a weighted distribution of [0.4, 0.3, 0.2, 0.1]. Then, we pick a phrase of the chosen length uniformly at random and repeat it.

**Pronoun Repetition.** We find pronouns to be commonly repeated in Indian languages. First, we accumulate a list of pronouns for each language. If any word in the fluent sentence appears in the pronoun list, then we repeat the pronoun with a predetermined probability.\(^4\)

\(^4\)These filler phrases are separately listed for all four Indian languages after consulting native speakers.

\(^5\)Here, we mean an n-gram of consecutive words regardless of their real phrasal structure.

\(^6\)This distribution was chosen only to signify that phrases of shorter length are more frequent than phrases of longer length.

\(^7\)More details about the probability with which a pronoun is chosen to be repeated is specified in our code.
3.3 Corrections

Corrections involve substitutions, deletions, or insertions of words from the reparandum section. Corrections may include the interregnum.

Partial Word. For this rule, we introduce partial words before long words with 12 or more characters. Firstly, we find the orthographic syllables of a long word using the Indic NLP Library (Kunchukuttan, 2020). Thereafter, we create the partial word by joining the first \( n \) syllables where \( n \) comes from a weighted distribution in which probability \( P(n) \) is proportional to \( \frac{1}{n^3} \).

Missing Syllables. For this rule, preceding a long word of 12 or more characters, we insert the same word but with one or more syllables missing. We first find the orthographic syllables of the long word. Then, we remove \( n \) contiguous syllables from the word (where \( n \) is sampled from a weighted distribution similar to what we used for phrase repetition) and add this reduced form of the word prior to the original long word.

Pronoun Correction. In this rule, a pronoun gets explicitly corrected. From the pronoun lists mentioned in Section 3.2, we create groups of similar types of pronouns (e.g., all first person pronouns are in one group). For each pronoun in the fluent sentence, we find its group and pick a different pronoun from the group to serve as its correction. We also (optionally) insert a frequent filler phrase before using the correct pronoun.

Synonym Correction. In this rule, we introduce a synonym of the word before the actual word, obtained using IndoWordNet (Bhattacharyya, 2010).

3.4 False Start

For the False Start disfluency, a sentence is aborted before it is completed, and a new idea or line of thought is introduced. To create false starts, we first randomly pick two different fluent sentences. Then, we split the first fluent sentence from a random position and we concatenate the first part of the split with the second fluent sentence.

4 Dataset Details

4.1 Disfluency Datasets for Indian Languages

Real Disfluent Data. We create real disfluent datasets in Bengali, Hindi, and Marathi by transcribing and annotating real disfluencies from conversations in the respective languages. For this purpose, we used publicly available Interviews in Bengali\(^\text{11}\), Hindi\(^\text{12}\), and Marathi\(^\text{13}\) from YouTube\(^\text{14}\). From these videos, we constructed three datasets containing 300, 150, and 250 disfluent and fluent parallel sentences in Bengali, Hindi, and Marathi, respectively.

Synthetic Disfluent Data. We also induce disfluencies in fluent text using our rule-based algorithm and create evaluation datasets for disfluency detection in Bengali, Hindi, Malayalam and Marathi. We start with fluent monolingual text from the PMIndia corpus\(^\text{15}\) (Haddow and Kirefu, 2020). We synthesize disfluent sentences using the rules outlined in Section 3. We ask language specialists in each of the four languages to manually pick sentences from the synthetic dataset that appear like natural disfluencies (and edit the disfluent sentences if needed). We picked utterances such that there is uniform coverage across disfluency types and there is no label imbalance. We used IndicNLP (Kunchukuttan, 2020) for normalization and tokenization, and we removed all punctuation marks. Table 1 shows detailed disfluency type counts for all four datasets. The test sets for Bengali, Hindi, Malayalam, and Marathi contain 500, 575, 575, and 420 sentences, respectively.

Each of the test sets is grouped into five categories: fillers, repetitions, corrections, false starts and fluent sentences. Fluent sentences are included as a control set to check whether the model is incorrectly detecting disfluencies in fluent sentences. We also include fluent sentences with reduplications which are a special category in Indian languages as mentioned below.

Reduplication. Reduplication is the act of repeating all or part of a word for emphasis or to...
We also present results for English to check how well MuRIL performs when compared with previously published results. Switchboard\(^\text{16}\) (Godfrey et al., 1992) in English is the most commonly used dataset for disfluency detection. Following the experimental settings in Wang et al. (2021), we split the Switchboard corpus such that the dev set consists of all sw_04[5-9]*.utt files, the test set consists of all sw_04[0-1]*.utt files, and the training set consists of all the remaining files. We do not include sentences without disfluencies in the training data, but do so in the dev, test set. Following Honnibal and Johnson (2014), we lowercase the text and remove all punctuation marks.

4.2 English Disfluency Data

We also present results for English to check how well MuRIL performs when compared with previously published results. Switchboard\(^\text{16}\) (Godfrey et al., 1992) in English is the most commonly used dataset for disfluency detection. Following the experimental settings in Wang et al. (2021), we split the Switchboard corpus such that the dev set consists of all sw_04[5-9]*.utt files, the test set consists of all sw_04[0-1]*.utt files, and the training set consists of all the remaining files. We do not include sentences without disfluencies in the training data, but do so in the dev, test set. Following Honnibal and Johnson (2014), we lowercase the text and remove all punctuation marks.

5 Experimental Setup

In this work, we use MuRIL (Khanuja et al., 2021) which is a BERT model (Devlin et al., 2019) pretrained on 16 Indian languages (including the four we consider) and English. MuRIL is pretrained using two language modeling objectives: Masked Language Modeling and Translation Language Modeling.

<table>
<thead>
<tr>
<th>Type</th>
<th>Bn</th>
<th>Hi</th>
<th>Ml</th>
<th>Mr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filler (3.1)</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>70</td>
</tr>
<tr>
<td>Word Repetition (3.2)</td>
<td>42</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Phrase Repetition (3.2)</td>
<td>42</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Pronoun Repetition (3.2)</td>
<td>41</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Partial Word (3.3)</td>
<td>66</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Missing Syllables (3.3)</td>
<td>34</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Pronoun Correction (3.3)</td>
<td>66</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Synonym Correction (3.3)</td>
<td>34</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>False Start (3.4)</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Fluent Sentences with Redpl</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>Normal Fluent Sentences</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>575</td>
<td>575</td>
<td>420</td>
</tr>
</tbody>
</table>

Table 1: Synthetic Dataset Statistics: Number of sentences of each disfluency type. Redpl: Reduplication, Bn: Bengali, Hi: Hindi, Ml: Malayalam, Mr: Marathi.

Convey a meaning. It is widely used in Indian languages; a few examples of reduplication in Hindi are shown in Table 2 (Montaut, 2009). In the context of disfluencies, we note that reduplications could be mistaken for a repetition disfluency type. Reduplications are intentional repetitions which are grammatically correct and should not be flagged as disfluencies. To check for this, we include fluent sentences with reduplication in our test set.

5.1 Using only English Disfluency Data

We finetune the pretrained MuRIL checkpoint on the English disfluency detection task where the goal is to correctly label each of the tokens as fluent or disfluent. We use the muril-base-cased checkpoint (having 236M parameters) from Hugging- Face\(^\text{18}\) for all our experiments. For each of the subword tokens identified by the MuRIL tokenizer, the model predicts its label as being 0 (fluent) or 1 (disfluent).

For disfluency correction, once we have disfluency labels for each subword, we use majority voting to determine whether a word is omitted or not. For a word, if the number of its subwords tagged as disfluent is greater than the number of subwords tagged as fluent, the word is deleted; else, it is retained.

5.2 Using Synthetically Generated Indian Language Data along with English Disfluency Data

We want to augment the Switchboard data with the synthetic Bengali data, but do not want the synthetic data to dominate the corpus.

**Evaluation Metrics.** We test the model on Bengali, Hindi, Malayalam, Marathi and English disfluency detection tasks. Similar to prior work on detecting English disfluencies (Wang et al., 2021), we compute precision, recall, and F1 scores using word-level labels. We also use a more ambitious metric, the exact match percentage, where the predicted fluent sentence is compared to the reference fluent sentence and checked for an exact match. We also show BLEU scores between the fluent text predictions and the reference fluent sentences, which are calculated using sacreBLEU (Post, 2018).

16https://catalog.ldc.upenn.edu/LDC97S62
17Orange color denotes reduplication
18https://huggingface.co/google/muril-base-cased
<table>
<thead>
<tr>
<th>Sentence</th>
<th>Transliteration</th>
<th>Gloss</th>
<th>Translation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>तुम कहा कहा गए</td>
<td>tuma kahaa kahaa gae</td>
<td>you where where went</td>
<td>where did you go</td>
<td>Reduplication of interrogative pronoun. Here the questioner expects a list of places in response.</td>
</tr>
<tr>
<td>खाते खाते मत बोलो</td>
<td>khaate khaate mata bolo</td>
<td>eating eating do not speak</td>
<td>do not speak while eating</td>
<td>Reduplication of verb</td>
</tr>
<tr>
<td>यह लो तुम्हारी चाय. गरम गरम है, पियो</td>
<td>yaha lo tumhari caaya. garama hai, piyo</td>
<td>this take your tea. hot hot is drink</td>
<td>Take your tea. It is nicely hot, drink it</td>
<td>Reduplication of adjective</td>
</tr>
<tr>
<td>बच्चो को एक एक टॉफ़ी दी</td>
<td>bacco ko eka eka taffii do</td>
<td>children to one one toffee give</td>
<td>give a toffee to each child</td>
<td>Reduplication of number.</td>
</tr>
</tbody>
</table>

Table 2: Examples of Reduplication in Hindi. Gloss: word-to-word English translation. “_” in the gloss suggests fertility which refers to one word mapping to multiple words in the other language.

Similarly, we construct synthetic data for the other two languages as well.

Next, we finetune the pretrained MuRIL checkpoint on the combined synthetic Bengali data and Switchboard data. The other experimental details are the same as described in Section 5.1. We evaluate this model on real disfluency detection data in Bengali, Hindi, and Marathi. We hypothesize that the performance of this model on the real Bengali disfluency detection dataset will be better than that of the model finetuned only on English data. This would indicate that our synthetically generated data contains disfluencies that mimic the ones seen in real speech.

5.3 Using Only Synthetically Generated Indian Language Data

We also finetune the pretrained MuRIL checkpoint only on the synthetic Bengali/Hindi/Marathi data. We apply the same synthetic data and experimental setup as discussed in Section 5.2.

6 Results & Analysis

This section presents the evaluation results and analyses the quantitative and qualitative performance of our models.

6.1 Performance on Real Disfluent Data

Table 3 shows a comparison of our model finetuned on only English data, and models finetuned on synthetic Bengali/Hindi/Marathi disfluent data (optionally) along with real English disfluent data.

We can see that MuRIL - En & Syn Bn (model finetuned using both synthetically generated Bengali data and the Switchboard corpus) outperforms MuRIL - En (finetuned only using the Switchboard corpus) by a significant margin of 19% in terms of exact matches, when evaluated on real Bengali disfluencies. Also, MuRIL - En & Syn Bn has high precision which leads to an increase of 4.67 F1 scores. Similarly, MuRIL - En & Syn Hi model outperforms MuRIL - En by a large margin of 19.92 F1 scores, when evaluated on real Hindi disfluencies. Both MuRIL - En & Syn Bn and MuRIL - En & Syn Mr also outperform MuRIL - En by 15.51 and 18.14 F1 scores, respectively. We observe similar trends on the real Marathi evaluation set as well.

All the models that were finetuned with additional synthetic data (irrespective of the language) nearly double the number of exact matches when evaluated on real Hindi/Marathi disfluencies, compared to the model trained with only English disfluent data. Our models trained only on synthetic disfluent data outperform the MuRIL - En model. We observe 1.17, 12.33, 17.48 F1 scores improvement over MuRIL - En model, when evaluated on real Bengali, Hindi and Marathi evaluation sets, respectively.

These results suggest that using synthetically generated disfluent sentences does enable transfer to real disfluent data and helps validate our synthetic data generation pipeline.

6.2 Disfluency Detection in Indian Languages

Table 4 presents a detailed account of the performance of MuRIL on the four manually-edited synthetic disfluency evaluation sets in Bengali, Hindi, Malayalam, and Marathi. It also provides a breakdown of performance across disfluency types. We report the exact match percentages, the BLEU scores and the F1 scores.

The overall F1 scores for Bengali, Hindi, Malayalam and Marathi are 73.14, 63.82, 67.12 and
Table 3: Performance on real Bengali, Hindi, and Marathi disfluent data. MuRIL - En: Finetuning MuRIL only on Switchboard data, MuRIL - En & Syn X (where X ∈ \{Bn, Hi, Mr\}): Finetuning MuRIL on Switchboard data and synthetic disfluency detection data in language X [Bn: Bengali, Hi: Hindi, Mr: Marathi] , MuRIL - Syn X (where X ∈ \{Bn, Hi, Mr\}): Finetuning MuRIL only on synthetic disfluent data in language X. We note that the BLEU scores between the original disfluent text and the fluent reference text for Bengali, Hindi, and Marathi are 62.0, 71.9 and 73.2, respectively.

<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Exact Match %</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bengali</td>
<td>MuRIL - En</td>
<td>81.13</td>
<td>65.60</td>
<td>72.54</td>
<td>28.33</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td>MuRIL - Syn Bn</td>
<td>91.79</td>
<td>61.58</td>
<td>73.71</td>
<td>43.00</td>
<td>81.2</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Bn</td>
<td>92.90</td>
<td>66.06</td>
<td>77.21</td>
<td>47.33</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Hi</td>
<td>89.90</td>
<td>51.03</td>
<td>65.11</td>
<td>35.67</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Mr</td>
<td>92.72</td>
<td>51.15</td>
<td>65.93</td>
<td>35.33</td>
<td>76.7</td>
</tr>
<tr>
<td>Hindi</td>
<td>MuRIL - En</td>
<td>67.81</td>
<td>62.20</td>
<td>64.89</td>
<td>36.67</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td>MuRIL - Syn Hi</td>
<td>83.18</td>
<td>72.05</td>
<td>77.22</td>
<td>54.00</td>
<td>89.5</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Bn</td>
<td>82.57</td>
<td>78.35</td>
<td>80.40</td>
<td>64.00</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Hi</td>
<td>84.98</td>
<td>84.65</td>
<td>84.81</td>
<td>66.00</td>
<td>93.7</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Mr</td>
<td>86.38</td>
<td>79.92</td>
<td>83.03</td>
<td>66.00</td>
<td>91.7</td>
</tr>
<tr>
<td>Marathi</td>
<td>MuRIL - En</td>
<td>57.78</td>
<td>54.93</td>
<td>56.32</td>
<td>26.40</td>
<td>83.8</td>
</tr>
<tr>
<td></td>
<td>MuRIL - Syn Mr</td>
<td>92.25</td>
<td>61.50</td>
<td>73.80</td>
<td>55.60</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Bn</td>
<td>82.77</td>
<td>68.78</td>
<td>75.13</td>
<td>56.80</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Hi</td>
<td>83.14</td>
<td>68.31</td>
<td>75.00</td>
<td>55.60</td>
<td>89.5</td>
</tr>
<tr>
<td></td>
<td>MuRIL - En &amp; Syn Mr</td>
<td>87.54</td>
<td>69.25</td>
<td>77.33</td>
<td>60.80</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Table 3: Performance on real Bengali, Hindi, and Marathi disfluent data. MuRIL - En: Finetuning MuRIL only on Switchboard data, MuRIL - En & Syn X (where X ∈ \{Bn, Hi, Mr\}): Finetuning MuRIL on Switchboard data and synthetic disfluency detection data in language X [Bn: Bengali, Hi: Hindi, Mr: Marathi], MuRIL - Syn X (where X ∈ \{Bn, Hi, Mr\}): Finetuning MuRIL only on synthetic disfluent data in language X. We note that the BLEU scores between the original disfluent text and the fluent reference text for Bengali, Hindi, and Marathi are 62.0, 71.9 and 73.2, respectively.

70.22, respectively. Interestingly, the model is able to do a reasonable job of disfluency detection even in the zero-shot setting with no access to labeled disfluent data in the target languages. The BLEU scores between the original disfluent text and the fluent reference text for Bengali, Hindi, Malayalam and Marathi are 83.2, 83.4, 78.3, 81.1, respectively. Comparing these scores to the BLEU scores obtained using the finetuned MuRIL (92.5, 90.9, 88.4 and 91.3) clearly shows that the model is effective in removing disfluencies.

6.3 Performance across Disfluency Types

Table 4 shows that our model is doing exceptionally well at detecting Repetitions in all the languages and our model shows the best performance in detecting Phrase Repetitions. In 64% of the Bengali sentences, our model did not tag a reduplication as disfluency, which is correct. This suggests that the model is learning the difference between word repetition and reduplication. Also, most of the fluent sentences are kept unchanged.

Even without any explicit supervision of the filler words specific to each language, we find that our model is sometimes able to accurately detect the fillers based on the context in which they appear. The high F1 score in Marathi for fillers (compared to the other three languages) can be attributed to the fact that the fillers exhibited a positional bias and mostly appeared at the start of the Marathi test sentences.

Our model performs fairly on detecting partial words. In comparison, the model does very well on detecting missing syllables and achieves more than 75 F1 scores in all four languages. Despite the complexity of the pronoun correction task, which could also involve optional editing terms, our model performs admirably and gets F1 scores of greater than 75 across all languages.

Detecting synonym correction correctly is a complex task and our model does not perform too well on this disfluency type. Our model yields the lowest F1 scores across all disfluency types on false starts. This is not very surprising because false starts are the hardest of disfluency types to detect (Shriberg, 1994). Sometimes there are ambiguities even in the gold standard utterances containing false starts. (Example 7 in Table 8 shows such an ambiguity.) Another reason could be that the Switchboard dataset does not contain many false start disfluencies.

6.4 Performance across Languages

From Table 4, we compare the performance of our model across languages for different disfluency types.

**Filler detection** is done best for Marathi. We observe that the model does not perform well in detecting Hindi fillers. Hindi differs from the other languages in that it uses fairly long phrases as fillers. For example, “क्या कहते है” means “what to say”. Thus, our model might find it challenging to catch these long filler phrases. We also observe that the model’s capability to detect a filler
When hesitations like “कारण” (meaning “because”), “विपय” (meaning “but”), “अंक” (meaning “and”) etc. as disfluencies which are part of the false start.

At times, the model detects part of the false start as being part of some other disfluency type, such as a repetition or a correction.

### 6.5 Ablation Study

Table 3 shows that combining synthetic Marathi disfluent data with English disfluent data increases F1 scores by 21 points on real Marathi disfluent data compared to a model trained solely on English data. Via an ablation study, we aim to check which subset of disfluency types in the synthetic dataset is most helpful. We consider seven combinations of disfluency types (including MuRIL - En20 and MuRIL - En & Syn Mr21) that we think are representative. We generate the same amount of synthetic data for all the combinations using the steps discussed in Section 5.2.

We present the results in Table 5. It is encouraging to see that all the models achieve a precision of nearly 90 when trained on synthetic data whereas Disf-onlyEng achieves a precision of only 57.78. In comparison to Disf-onlyEng, Disf-1 (synthetic data containing only fillers) has a lower F1 score, which can be attributed to the very low recall value. Disf-1234 improves F1 score by 8.40, which suggests that repetitions help. It is interesting to see that Disf-15678 improves F1 score by 20.09. This implies that corrections by themselves are of sig-

---

Table 4: Performance on manually-edited synthetic Bengali, Hindi, Malayalam, and Marathi disfluency datasets. We use MuRIL - En model for the evaluation. M: Exact Match Percentage, B: BLEU Score, F1: F1 Score. We note here that the BLEU scores between the original disfluent text and the fluent reference text for Bengali, Hindi, Malayalam and Marathi are 83.2, 83.4, 78.3, 81.1, respectively.
We propose the use of a pretrained multilingual model MuRIL for zero-shot disfluency detection in Indian languages. We evaluate our model on Bengali, English, Hindi, Malayalam, and Marathi disfluency detection tasks. We also show that synthetically generated Bengali/Hindi/Marathi disfluency detection data using simple rules, when combined with real English disfluency data during finetuning, helps improve F1 scores on real Bengali/Hindi/Marathi disfluencies. Our overall results support the claim that it is possible to do cross-lingual transfer of disfluency detection without any labeled data in the target languages. For future work, we intend to evaluate the model on more diverse disfluencies.

### Acknowledgements

The authors thank the anonymous reviewers for their constructive suggestions that helped improve the draft. This work was done as part of the Bahubhasak Pilot Project on Speech to Speech Machine Translation under the umbrella of National Language Technology Mission of Ministry of Electronics and IT, Govt. of India.

### References


---

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
<th>Exact Match %</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disf-onlyEng</td>
<td>57.78</td>
<td>54.93</td>
<td>56.32</td>
<td>26.40</td>
<td>83.8</td>
</tr>
<tr>
<td>Disf-1</td>
<td>90.97</td>
<td>30.75</td>
<td>45.96</td>
<td>23.60</td>
<td>79.9</td>
</tr>
<tr>
<td>Disf-1234</td>
<td><strong>93.36</strong></td>
<td>49.53</td>
<td>64.72</td>
<td>45.60</td>
<td>84.8</td>
</tr>
<tr>
<td>Disf-15678</td>
<td>89.06</td>
<td>66.90</td>
<td>76.41</td>
<td>56.80</td>
<td>89.4</td>
</tr>
<tr>
<td>Disf-137</td>
<td>89.78</td>
<td>47.42</td>
<td>62.06</td>
<td>38.80</td>
<td>83.6</td>
</tr>
<tr>
<td>Disf-1357</td>
<td>89.56</td>
<td>62.44</td>
<td>73.58</td>
<td>54.00</td>
<td>88.2</td>
</tr>
<tr>
<td>Disf-all</td>
<td>87.54</td>
<td><strong>69.25</strong></td>
<td><strong>77.23</strong></td>
<td><strong>60.80</strong></td>
<td><strong>90.0</strong></td>
</tr>
</tbody>
</table>

Table 5: Performance of models trained on English and additional synthetic Marathi disfluent data containing a subset of disfluency types (except Disf-onlyEng which is trained only on English data). The numbers in the model name indicate which disfluency types were present in the synthetic data during finetuning (e.g., Disf-137 was trained on synthetic Marathi disfluent data containing only filler, phrase repetition, and pronoun correction). Mapping of number to disfluency types — 1: filler, 2: word repetition, 3: phrase repetition, 4: pronoun repetition, 5: partial word, 6: missing syllables, 7: pronoun correction, 8: synonym correction. We note here that MuRIL - En is the same as Disf-onlyEng and MuRIL - En & Syn Mr is the same as Disf-all.

Table 6: Performance on Switchboard disfluency detection test set.

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>95.22</td>
<td>92.08</td>
<td>93.62</td>
<td>98.00</td>
</tr>
</tbody>
</table>

Table 6: Performance on Switchboard disfluency detection test set.


### A Examples: Synthetic Disfluency Generation Rules

Table 7 presents a brief overview and example of each type of disfluencies covered in Section 3.

#### 22 Purple color denotes disfluencies

### B Analysis: Inconsistency in Predicted Tags of the Subwords of a Word

Since our model works at the subword level, it could be possible that the subwords of a particular word get different tags (*disfluent/fluent*) as the prediction. We call this *inconsistency in tagging*.

According to our findings, only 0.02 % of subwords in the Switchboard test set were marked inconsistently. In our Bengali, Hindi, Malayalam, and Marathi test sets, we found inconsistency in just 0.21 %, 0.06 %, 0.38 %, and 0.03 % subwords. These findings indicate that our model is quite likely to consistently predict the tags across the subwords for any word.

### C Qualitative Analysis

We analyze the performance of our base model (trained only on English data) on relatively difficult disfluencies, namely, *false starts* and *corrections*. We also analyse one example of a fluent sentence and we present these examples of potential interest in Table 8.

#### 23 Cyan color denotes that it is unclear whether the words are intended in the sentence
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Example</th>
<th>Transliteration</th>
<th>Gloss</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filler (3.1)</td>
<td>Insert frequent filler phrases in a sentence</td>
<td>এই মাত্রের কৃত্তিত্ব এবং চর্চার প্রথা থেকে যেই অভিকাশ বাঁকা হয়।</td>
<td>samikiShYAya dekhaAYa YAYa, ei mAne teRRi-tiAYa evaM chaturtha shreiN1 thekei adhihikAMsha ChAtri skulaChuTa haiYa</td>
<td>still</td>
<td>According to the survey, this means that most girl students drop out of school from third and fourth grade.</td>
</tr>
<tr>
<td>Word Repetition</td>
<td>Repeat a word unnecessarily</td>
<td>এখনও এখনও হতে আমারের বাইরে থেকে সত্ত্বে থেকে।</td>
<td>ekhanao ekhano haYato anakeke vAire theke jala Anate Yete haiYa</td>
<td>still</td>
<td>Many may still have to fetch water from outside.</td>
</tr>
<tr>
<td>Phrase Repetition</td>
<td>Repeat a phrase unnecessarily</td>
<td>অধিকতর হতে আমারের বাইরে থেকে সত্ত্বে থেকে।</td>
<td>arthAt, hatAshAra Avadeo vA.NchAra AshA jAgAnora sAmArthaYa ei samAjera rYaYche</td>
<td>in other</td>
<td>In other words, this society has the ability to inspire hope to survive in the face of despair.</td>
</tr>
<tr>
<td>Pronoun Repetition</td>
<td>Repeat a pronoun unnecessarily</td>
<td>অর্থাৎ, হতে আমারের বাইরে থেকে সত্ত্বে থেকে।</td>
<td>AmarA AmarA roYAn,DAvAsIlke eckShetre nAnAb-hAve sAhAYYa karate pAri</td>
<td>we</td>
<td>We we to Rwandans in this case many ways help to do can</td>
</tr>
<tr>
<td>Partial Word (3.3)</td>
<td>Use part of word before the actual word</td>
<td>এর ফলে, নায়ক অক্ষরের সমস্যা সমাধান করতে পারি।</td>
<td>era phale, naYa.DA anchale AgAnl vaCharagulite janasaMkhYAo ulleka ullekaYo-gYabhAve vRRiDhii pAve</td>
<td>for this</td>
<td>As a result, the population of the Noida region will also increase significantly in the coming years.</td>
</tr>
<tr>
<td>Missing Syllables</td>
<td>Missed a few syllables from the middle of a word; therefore, it is followed by the entire word.</td>
<td>আটিয়ান হটিনিরসের ইন্টেলিজেন্সের সৃত্রথিত দৃষ্টি করছ সমস্যার সমাধান করতে পারি।</td>
<td>ArtTiphishiYaAla inTelinsera in-Telijenseru sUl-tra sharei vahu samasYara samAdhAha AmarA karate pAri</td>
<td>Artificial</td>
<td>We can solve many problems with the help of Artificial Intelligence.</td>
</tr>
<tr>
<td>Pronoun Correction</td>
<td>Use an incorrect pronoun, then an optional edit phrase, then the proper pronoun.</td>
<td>কাছে এদেরকে আমারের বাইরে থেকে পারি।</td>
<td>kachChe ederake nA mAnE eke bhu NgA vale</td>
<td>in Kachchh</td>
<td>We no I mean it is called bhunga in Kachchh.</td>
</tr>
<tr>
<td>Synonym Correction</td>
<td>Use of imprecise synonym before the actual word</td>
<td>এই মাত্রের কৃত্তিত্ব এবং চর্চার প্রথা থেকে যেই অভিকাশ বাঁকা হয়।</td>
<td>kon prakalpa kave sheSha have, sei samaYa nirdeshita nirdiShTa kare deoYaiA haYaeC</td>
<td>which project when complete will be, that time directed specified has been done</td>
<td>When the project will be completed, the time has been directly specified.</td>
</tr>
<tr>
<td>False Start (3.4)</td>
<td>Begin a sentence, then abruptly end it and begin a new sentence.</td>
<td>ইতঃমধ্যেই জাতো আমারের বাইরে থেকে যেই অভিকাশ বাঁকা হয়।</td>
<td>itimadhiYe rAjYe Aja desha unnaYanera natuna uchchetA atikrama karaChe</td>
<td>already</td>
<td>Already in the state today country of development new height exceed doing.</td>
</tr>
</tbody>
</table>

Table 7: Different types of disfluencies in our synthetic dataset
<table>
<thead>
<tr>
<th>Disfluent Sentence</th>
<th>Transliteration</th>
<th>Gloss</th>
<th>Translation</th>
<th>Model Output</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>কের কাজ করব</td>
<td>BT: Asde Ami nk</td>
<td>Ami ekadamai vujhe pxtChA nA kI kara uchita</td>
<td>E: Actually I no have no idea what to do</td>
<td>ভাবিবেন পঠনের ফলে তুমি সেই কাজ করবেন</td>
<td>Our model is able to detect the correction.</td>
</tr>
<tr>
<td>ছয় মাস ধেরেকাম্পানি</td>
<td>BT: met nahlM matak laba hamARa atha aMdarUl tuba para hameshA uMyama-pUvaka buna_paja bai</td>
<td>E: My no I mean our past has always been abominously woven in</td>
<td>E: Our model is able to detect the disfluency.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>আত্মার মােক ঈশ্বরকৃত</td>
<td>ET: Chaya MAsa</td>
<td>অল্প ভাবিবেন চার বছর ধের এটা ভাবেল</td>
<td>E: The model is able to detect the correction.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ওই কাজটা হৈ</td>
<td>ET: meNi Chaya MAsa</td>
<td>অল্প ভাবিবেন চার বছর ধের এটা ভাবেল</td>
<td>E: The model is able to detect the correction.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Qualitative analysis of our model predictions. BT: Bengali Transliteration, BG: Bengali Gloss, HG: Hindi Gloss, MrT: Marathi Transliteration, MrG: Marathi Gloss, E: English Translation.