Towards Disfluency Annotated Corpora for Indian Languages

Chayan Kochar, Vandan Mujadia, Pruthwik Mishra, Dipti Misra Sharma
LTRC - International Institute of Information Technology Hyderabad
{chayan.kochar, vandan.mu, pruthwik.mishra}@research.iiit.ac.in, dipti@iiit.ac.in

Abstract

In the natural course of spoken language, individuals often engage in thinking and self-correction during speech production. These instances of interruption or correction are commonly referred to as disfluencies. When preparing data for subsequent downstream NLP tasks, these linguistic elements can be systematically removed, or handled as required, to enhance data quality. In this study, we present a comprehensive research on disfluencies in Indian languages. Our approach involves not only annotating real-world conversation transcripts but also conducting a detailed analysis of linguistic nuances inherent to Indian languages that are necessary to consider during annotation. Additionally, we introduce a robust algorithm for the synthetic generation of disfluent data. This algorithm aims to facilitate more effective model training for the identification of disfluencies in real-world conversations, thereby contributing to the advancement of disfluency research in Indian languages.

Keywords: disfluency, annotation guidelines, synthetic augmentation, Indian languages

1. Introduction

Natural speech has its own uniqueness. Written text tends to be very fluent and can be readily used for NLP tasks after preprocessing. In contrast, people often think and speak simultaneously during a discussion when speaking naturally. Individuals often exhibit a reflexive tendency to rectify errors upon recognizing inaccuracies in their speech. This can involve editing, reformulating, or even starting over from scratch. This is a normal, intuitive process that seamlessly gets mixed in spontaneous conversational interactions. Thus natural speech often exhibits such interruptions and disruptions known as disfluencies (Shriberg, 1994).

Disfluencies can be classified into mainly 5 categories: filler words, pet phrases, repetitions, repair and false starts. Though there are other naming conventions or groups which may overlap with the ones mentioned, but there is a distinct characteristic to every disfluent utterance. Every disfluent utterance or a phrase comprises of a ***reparandum***, usually followed by a verbal cue, the *interruption point*, an optional ***edit term***, and finally the optional ***alteration*** (Shriberg, 1994; Heeman and Allen, 1999). The alteration is what an ideal fluent text would have been, replacing the reparandum and editing terms.

These phenomena encompassing hesitations, repeats, corrections etc which are so abundant yet unnoticeable, is what makes the problem so interesting. These disfluent terms can be eliminated because they do not add to the semantics of the sentence, thus producing a noise free data ready to feed to the machines.

This very aspect makes disfluency correction a very crucial factor for any other NLP downstream tasks like MT (Rao et al., 2007; Wang et al., 2010), question answering (Gupta et al., 2021) etc. If the fundamental tasks like these are jeopardized, then all other tasks following them would yield poor output as well. To get this done, a robust set of annotation guidelines is paramount for ensuring the quality, consistency, and reliability of annotated data in any research endeavor, particularly in the field of Natural Language Processing (NLP). A set of detailed annotation guidelines would bring in consistency, reduced ambiguity, scalability and most importantly cross dataset compatibility due to abundance of linguistic features which are common in Indian Languages.

India has a rich linguistic diversity, with about 1369 rationalized mother tongues and numerous more under resourced languages [1]. Given the vast array of linguistic nuances and variations present in India, any NLP-related problem-solving approach must account for this diversity to ensure its effectiveness and applicability within the Indian context. Therefore, it is imperative to prioritize the development of NLP technologies tailored to the specific linguistic landscape of India, facilitating broader accessibility and utility for its diverse population. In light of the lack of good amount of labelled data for Indian languages, the concept of synthetic augmentation becomes much more relevant. Due to the newly created dataset's wide range of variations and scenarios, it not only tackles the issues of data scarcity and class imbalance but also improves model generalisation. Research suggests that this indeed helps in overall performance of the model. (Passali et al., 2022; Kundu et al., 2022)

Extensive research has been conducted on disfluencies (Colman and Healey, 2011; Shriberg,

[1] Census 2011

WILDRE-7 Workshop, pages 1–10
25 May, 2024.

1994) along with work on identification and/or removal of such disfluencies (Wang et al., 2020). Though most of the work have focussed on English, and not much work has been contributed when it comes to disfluencies in Indian Languages. This lack of research for Indian context can be attributed to the scarcity of labeled data and standardized annotation guidelines specific to Indian languages. In this research, we aim to bridge that very gap along with providing a robust algorithm for synthetic generation of required data.

An example sentence showcasing the importance of disfluency handling for MT (Hindi -> English):

तो उधर आपको सर ने क्या कहा था ? क्यों बोला था उनको ? बोला, क्या प्रॉब्लम है उनको ?

Corresponding Translation to English: So what did Sir tell you there? Why did you tell him? Said, What problem does he have?

Translation after removing disfluency: So what did Sir tell you there? What problem does he have?

The text in blue indicates the alteration, while the text in red indicates the reparandum (category: repair). After the reparandum is removed from the original text, the translation quality becomes better.

2. Related Work

Previous research has primarily focused on speech and spoken disfluency, with limited attention given to textual disfluencies. Moreover, research specifically addressing disfluencies in Indian languages is scarce. It is important to note that there is a notable absence of standardized annotation guidelines tailored for annotating disfluencies in Indian languages. Therefore, this is an aspect that we propose to address through our work.

A very efficient solution is available for generating disfluent data in English (Passali et al., 2022). However, when considering Indian languages, it is not feasible to directly apply similar algorithms for the reasons outlined above. (Bhat et al., 2023b) investigated a dataset for disfluency correction, though their focus was solely on Hindi among the Indian languages.

For Indian Languages, a zero shot detection of disfluencies along with synthetic data generation was shown to be very useful (Kundu et al., 2022). This shows us the reason why such synthetic augmentation can be so crucial. Due to the newly created dataset's wide range of variations and scenarios, it not only tackles the issues of data scarcity and class imbalance but also improves model generalisation. Additionally, research has demonstrated that adversarial training with actual data but a significant reliance on synthetic data also improves score. (Bhat et al., 2023a). Our analysis of disfluencies exhibits a finer granularity in annotation and construction, which extends to Indian languages and tries to surpass previous research efforts in this domain.

Disfluencies are perfectly natural, and do not sound wrong to the human ear. When we talk about disfluency correction, we primarily try to make the machine understand better. Ultimately, in the bigger picture of speech-to-speech machine translation, we would want the output to be as human-like as possible. Since disfluency in one language does not necessarily map one-to-one with another language, hence it is indispensable to know the complicacies regarding disfluencies in both source and target language.

The primary focus areas in this study are:

- Appropriate annotation guidelines that consider the subtleties of Indian languages
- Synthetic generation of such disfluencies using an algorithm that tries to improve on previous works
- Characteristics of Indian languages which might appear very similar, but are different to disfluencies
- How code-mixed data plays a role

Here we work on 6 Indian Languages namely: Hindi, Bengali, Marathi, Telugu, Kannada and Tamil.

3. Data

We used simulated conversations in authentic contexts for our investigation. This was obtained by us from the IIT Madras SPRING lab [2],- who had acquired this data from vendors on a payment basis followed by thorough quality check on the transcriptions. The dataset included both monologues and conversations between two to four persons.

Since monologues are usually prepared or practiced speeches, people frequently have the chance to plan and organise their speech beforehand, reducing the likelihood of disfluencies like pauses, hesitations, or self-corrections. Furthermore, the lack of instant input from listeners lessens the necessity for spontaneous alterations or changes during monologues. Thus we focussed on the natural conversational audios. We manually filtered

[2] https://asr.iitm.ac.in/dataset

Language	Synthetic	PMIndia	Real data
Hindi	7	1	5.5
Bengali	7	1	5.5
Marathi	7	1	2
Telugu	5	1	2.5
Kannada	7	1	3
Tamil	7	1	8

Table 1: Full distribution of data (in hours)

out those non-monologue audios which appeared to have good amount of disfluencies. We acquired the data for Hindi, Marathi, Bengali, Kannada and Tamil as mentioned above, whereas for Telugu, we collected the data using conversations from YouTube videos with creative commons licence. While annotating on the acquired transcripts, if any instance arose regarding incorrect transcription, we first rectify the transcript before proceeding with the annotation.
We also used approximately 1 hr[3] of data from the PMIndia (Haddow and Kirefu, 2020) corpus for each language, to which we synthetically added disfluency. This allowed to make our dataset more diverse.

The Table 1 shows the distribution and size of data set for each language(numbers).

3.1. Tagset Considered

The tags considered for annotating the data include:

- **Pet_r** : marks the reparandum under the category of pet_phrases
- **Filler_r** : marks the reparandum under the category of filler words/pauses.
- **Edit_r** : marks the edit terms, also called the interregnums (Kundu et al., 2022)
- **Repeat_r** : marks the reparandum under the category of repetition
- **Repair_r** : marks the reparandum under the category of repair
- **False_r** : marks the reparandum under the category of false start
- **Alteration**: marks the alteration where required.

[3]We approximate 6500-7000 words to be present in one hour of speech

3.2. Annotation Guidelines

In contrast to English, the datasets comprising six Indian languages lack comprehensive guidelines regarding their behavior concerning disfluency.
Hence we followed a holistic approach for identifying the instances which count as disfluency, along with identifying other minute details which need to be given special attention to while dealing with Indian languages.
A pivotal concept reiterated throughout is the variability of words or phrases that may exhibit disfluency in certain contexts but not in others. This variability hinges on whether the word or phrase carries semantic significance in the given context.

The following examples use red text to indicate reparandum, green text to indicate editing terms, and blue text to indicate alteration. Unless specifically mentioned, the non-English examples are in Hindi. For all the non-English text, its corresponding transliteration is present under the respective texts.

3.2.1. Filled Pauses/Filler Words

This category encompass the phenomena when speakers tend to use certain sounds like 'uh', 'uhmm' in between their utterances. These do not carry any meaning, and in most cases just a sign of the speaker thinking and speaking simultaneously. Important exception: the cases of interjections and discourse markers. There are cases where certain filler words are used as meaningful interjections/discourse markers. In such cases, they should not be marked as disfluency.

Examples:

- Hindi: मैं अ कल तक अ पहुंच जाउँगा
  mai uh kal tak uh pohoch jaunga
- Tamil: அவளுக்கு ஒரு ம்ம் குறுஞ்செய்தி அனுப்பு
  Avaḷukku oru m'm kurunceyti anuppu

3.2.2. Pet Phrases

Many speakers use particular terms rather frequently, even in situations where their semantic contribution is negligible. These terms are called "pet phrases", and they can include discourse markers and common interjections. Moreover, these catchphrases are unique to each speaker, and there is no set list of terms that they can use as their pet phrases.

Examples:

- Hindi: मतलब यह बात मतलब एक मेम बोले थे
  matlab yah baat matlab ek ma'am bole the

- Marathi: आपन काल ते खाल्लं नं, ते आपलं खरबूज, ते खूप छान होतं.
  aapan kaal te khaalla na, te aapla kharbuj, te khup chaan hota

3.2.3. Repetitions

These are the simple cases when speaker repeats certain words/phrases in continuation. We need to be cautious when dealing with the concept of reduplication and emphasis in Indian Languages (Section 3.3). Those cases should not be marked as disfluency.

Examples:

- Hindi: एक नार्मल डॉक्टर का डॉक्टर का उस दिन आना जरूरत था ।
  ek normal doctor ka doctor ka us din aana jaroorat tha

- Bengali: আমি বাড়িতে পৌঁছে এটি সমাধান এটি সমাধান করার চেষ্টা করব
  aami barite paunche eti samaadhaana eti samaadhaana korar chesta korbo

3.2.4. Repair

There are many instances where the speaker utters words and phrases, but then realizes his mistake, and corrects it. The part which he uttered by mistake is part of the reparandum, and the alteration contains the part to be replaced with. Importantly, the topic remains the same. There are cases of emphasis, code mixing, echo words, abrupt endings, phrase insertion(gaps) which should not be confused with the phenomena of repair. Section 3.3 contains details for all such cases.

Examples:

- Hindi: वी थिंक दॅट मतलब हम बस एक ही चीज सोचते कि इन्हें बेस्ट पॉसिबल ट्रीटमेंट मिले, चाहे वो कैसे भी मिले.
  we think that matlab hum bas ek hi cheez sochte ki inhe best possible treatment mile, chahe wo kaise bhi mile

- Telugu: నేను రేపు అహ ఎల్లుండి వెళ్తున్న,
  nenu repu aha ellundi veltunna

- Bengali: আমার ফ্লাইট আগামীকাল সকাল ৭টায় বিকাল ৭টায়.
  aamaar flight aagamikaal shokal shaattaay bikal shaattaay

3.2.5. False Start

In this phenomena, the speaker abandons his utterance midway through and starts with another utterance with a different topic. We simply note the editing and reparandum terms since false starts indicate that the speaker is beginning over. This is due to the lack of clarity regarding the precise alteration that would be made — either the entire sentence or just a portion of it. As a result, we do not mark any alterations to avoid any ambiguity.

Examples:

- Hindi: मैंने अ.. ऑलरेडी मेरा इन्शुरन्स इनिशिएट हो चुका था ।
  maine uhh.. already mera insurance initiate ho chuka tha

- Marathi: कालचा एपिसोड तर... अरे आपल्याला दिवाळी च्या सुट्ट्या कधी आहेत ?
  kaalcha episode tar... arey aaplyala diwali chya suttya kadhi aahet?

3.2.6. Edit term

These are the lexical cues which indicate the end of reparandum and start of alteration. It can be filler words/pet phrases, or some words distinctively carrying the meaning of 'apology the unintended utterance(reparandum)' which are exclusively considered as edit terms like "sorry", "i mean" in English. These are marked in the above mentioned examples in green color.

3.3. Corner cases while annotating disfluencies

All the below instances are not to be considered as disfluencies, except code mixing in certain scenarios, as explained.

3.3.1. Code Mixing

Many instances involving code mixing, may or may not be part of disfluencies. This can be identified as following.

Cases when Code Mixing is NOT disfluency:

- **Simple Code mixing**: This is when we replace the words/phrases of one language with another, without interrupting the flow of speech.

  Example: I was going to reach my home कि माँ का फोन आ गया
  I was going to reach my home ki maa ka phone aa gaya

- **Emphasis**: There will be instances where a speaker deliberately utters a sentence in one language and repeats in another, to emphasise

its importance. Usually this kind of emphasis occurs when the whole clause/sentence is repeated in another language. The most helpful cue to detect emphasis of such kind is from the audio.

Example: I will do the work by tomorrow मैं कल तक काम कर लूंगा ।
I will do the work by tomorrow main kal tak kaam kar lunga

- **Situational Code mix**: These are instances when a speaker deliberately uses another language and repeats what he said, to make sure the other speaker is following him.

Cases when Code Mixing leads to Disfluency:

- When the speaker starts in one language, abandons it midway and utters the same sentence in another language. - this will be an instance of *repair* type of disfluency

  Example: आइ विल मैं कल तक काम कर लूंगा
  I will main kal tak kaam kar lunga

- When the speaker starts in one language, abandons it and utters a new sentence with topic change in another language - this will be an instance of *false start* type of disfluency.

  Example: Her name मैं वहाँ दस बजे तक पहुँच जाऊँगा
  her name main wahan das baje tak pahunch jaunga

Both these techniques were also applied while generating synthetic disfluencies.

3.3.2. Reduplication

This is a phenomenon widely present in Indian Languages. The speaker deliberately repeats certain words to convey some meaning/emphasize.

Example: ऐसे छोटे छोटे बातों पे वो चिल्लाने लगते थे ।
aese chote chote baaton pe wo chillane lagte the

3.3.3. Echo Words

This is a similar phenomenon to reduplication, but the words are not exactly copied, rather they sound/rhyme similar.

Example: मतलब सिस्टर–विस्टर से पूछने की कोशिश करते हैं
matlab sister-wister se puchne ki koshish karte hain

3.3.4. Emphasis

- **By Repetition**: Frequently, speakers intentionally repeat a word or phrase to emphasize a point or convey specific meaning.

  Example: तो जनरल वार्ड में पहले मेडिसिन्स ही चल रहा था । चल रहा था ,चल रहा था ।
  toh general ward me pehle medicines hi chal raha tha chal raha tha chal raha tha.

  In this instance, despite the repetition of words or phrases, nothing has been labeled as reparandum or alteration. This repetition is intentional on the part of the speaker telling about the continuous process of administering medicines.

- **Numbers**: Speakers often emphasise on what they want to convey by simply repeating the numbers or say the same thing in different languages(code mixed). Such cases are not to be considered disfluency.

  Example: बहोत कॉस्टली है, पैंतीस से चालीस हजार थर्टी फाईव्ह टू फोर्टी थाउजंड पर डे, लग रहे है
  bohot costly hai, pentees se chaalees hazaar thiry five to forty thousand per day lag rahe hain

  In this example, there would not be any disfluency - neither repair nor a code mixed repeat. By repeating in different languages, the speaker simply emphasizes the huge sum the figure represents.

- **Specificity**: Speakers often tend to specify about what they uttered, giving specificity to certain words/nouns.

  Example: मुझे क्रिकेट खलेने के लिए बॉल लाल बॉल चाहिये.
  mujhe cricket khelne ke liye ball laal ball chahiye

3.3.5. Abrupt Endings

There is also the presence of 'abrupt endings', where the speaker altogether leaves some useful meaningful utterance midway and starts other utterance. In such cases, they are not disfluency.

Consider the text:
तो उसे अगर इधर दर्द देता है तो सिस्टर को बुला कर थोड़ा मतलब ये प्रॉब्लम नहीं है । तो वो सिस्टर्स ध्यान से देख लेते हैं।
**toh use agar idhar dard deta hai to sister ko bula kar thoda matlab* ye problem nahi hai. toh wo sisters dhyan se dekh lete hai .*

Here, following the portion enclosed in asterisks, the speaker discontinues and initiates a fresh expression or, alternatively, substitutes the entire phrase with 'ये'. One cannot consider this as disfluency since all parts of the asterisked utterance is important and conveys some information. Removing them would result in loss of information - which is not what disfluency represents.

3.3.6. Different Speakers

It is essential to note that when marking any utterance as disfluent, we must ensure that the suspected disfluent utterances originate from a single speaker. Thus the cases of *'echoic utterance'* or *'echoic questioning'*, indicated by a speaker repeating or echoing the listener's response immediately - should not be termed as disfluency.

Likewise, there are scenarios in which a speaker is interrupted mid-sentence by another speaker, resulting in an apparent disruption in the conversation flow. Nevertheless, such instances should not be labeled as disfluencies.

Addressing these nuances poses a significant challenge in disfluency identification tasks, particularly in the context of real conversations.

3.3.7. Phrase Insertion or Gaps

Indeed, it is common during speech for individuals to interject additional information abruptly to provide better context before resuming their original train of thought. Such instances should not be termed as disfluencies.

Consider the following example:
तो उसके बाद वो बोले मेरे से बात हुआ था कि ऐसा ऐसा है, तो आपको इमिडिएटली पैसा रिलिज करना पडेगा ।
**toh uske baad wo bole* mere se baat hua tha ki esa esa hai to aapko immediately paisa release karna padega*

If we observe, the portion enclosed in asterisks is where the speaker moved off from his speech, got some additional information as underlined, and then continued from where he left off. Thus these types of instances are special to spoken language, but not any disfluency.

Figure 1 shows a glimpse of the in-house developed tool to perform the annotations. We used our own tool so that we can easily customize the tags as well as carry out simultaneous editing of subtitles with respect to the audio playback.

4. Inter Annotator Agreement

Two annotators worked on each language for the task of annotating disfluencies. Thus correspondingly, the inter annotator agreement was done on 1 hr of data of each language. To calculate the IAA, Cohen's kappa coefficient was used - giving a score of 82-92% for the disfluency annotation on each language. Table 2 shows the kappa scores for the annotations done for disfluencies on Indian Languages.

Upon thorough analysis, it was found that pet phrases and filler words had the highest degree of agreement among annotators. On the other hand, there were some differences in the repair and false start annotation cases. It is crucial to remember that these differences did not always imply different annotations for the reparandum as a whole. Rather, disagreements primarily centered around the span of words marked for the reparandum and alteration. Considering the simplicity of filler words compared to the complexity and intricacies of repair annotations, this result is in line with our predictions.

Language	Score
Hindi	92
Bengali	89
Marathi	83
Telugu	86
Kannada	82
Tamil	85

Table 2: Kappa metric scores for annotation in each language

Thus in this task of disfluency identification, about 85-86% of inter-annotator agreement on average was reached. This level of agreement shows that we have a strong and dependable method for spotting and categorizing speech interruptions in different language situations.

5. Synthetic Data

When synthetically augmenting data with disfluencies, we have to keep in mind to make the disfluent data look as natural as possible.
To achieve the same, we synthetically generated the five categories along with paying attention to the different kinds possible within each category.

5.1. Filler words, Pet Phrases

For each language we had collected a list of commonly used filler words. Then for the given sentence, a random position is chosen, except the last position, and a random filler word from the list is concatenated at that position.
It is significant to note that when a speaker utters

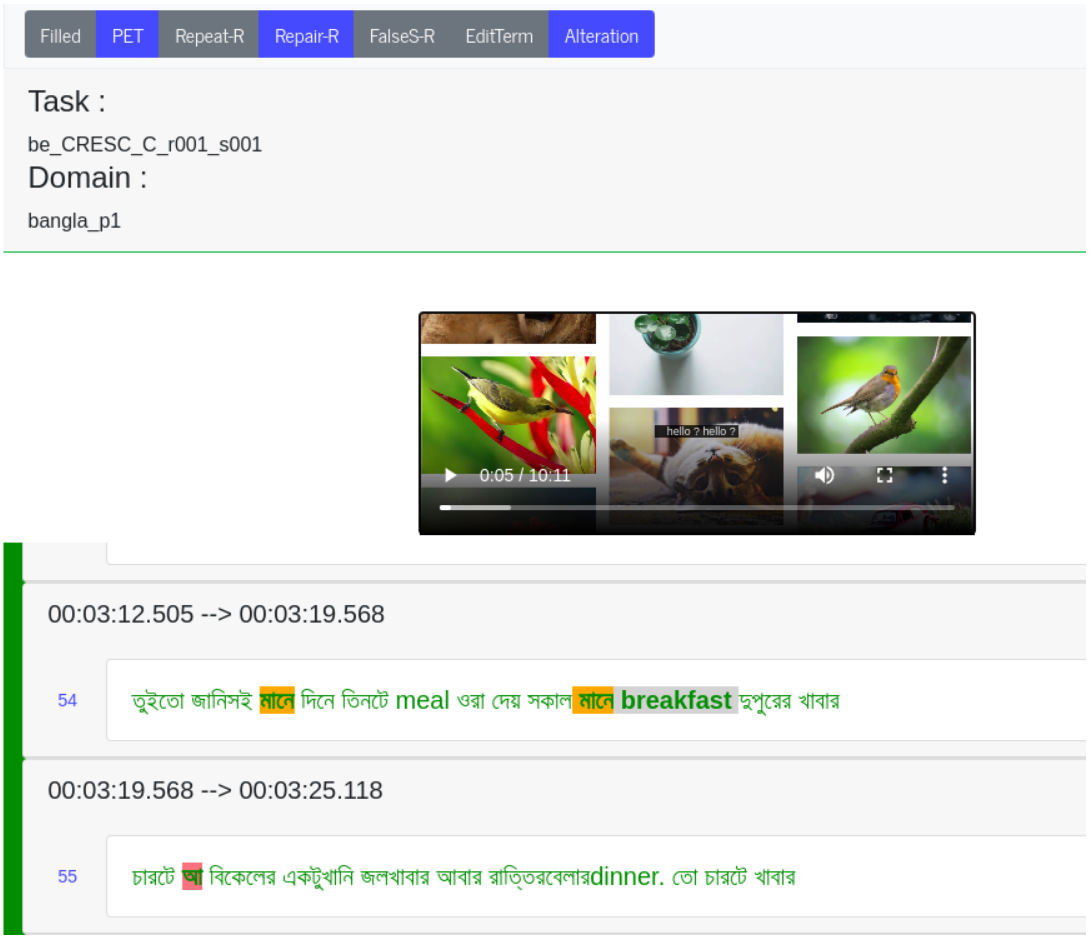

Figure 1: Tool used for Annotation(in-house developed tool). This is just an illustrative image showing the audio playback along with the corresponding subtitles (here in Bengali). The annotator can select the text and click on the appropriate category from the tabs present at the top. Each category will have its unique highlight color as well. The categories are namely: Filled Pause, Pet Phrase, reparandums of Repair, Repeat and False Start, Edit Terms and Alteration.

a filler word, he does not do it just once. Since fillers are a sign that the speaker is thinking and speaking, filler words usually occur more than once in an utterance or a sentence.
Hence following the same notion, if we are injecting a filler word to a sentence, we take into the account the length of the sentence, a probability measure 'p', and the max filler words that a sentence can accommodate - which we capped at 4. Thus using this methodology, we augmented filler words in a given sentence.
What distinguishes a pet phrase from a filler word in this methodology is the notion that pet phrases are unique to an individual. Therefore, if the speaker's identity is known in advance, the same pet phrase previously used by that individual is employed with higher likelihood.

5.2. Repetition

To add Repetition type of disfluency, we follow a similar approach as mentioned in (Kundu et al., 2022).

- **Word Repetition** : To implement this we choose a word randomly and repeat it.
- **Phrase Repetition** : We repeat an n-gram of two to five words. We initially use a weighted distribution of [0.4, 0.3, 0.2, 0.1] to randomly select a length from [2, 3, 4, 5].

5.3. Repair

The phenomenon of repair has many nuances if we observe carefully.

- **Partial Word**: This represents the concept wherein a speaker partially utters a word, then properly utters it afterwards - closely related to stammering.

  Attention was paid on how and where such disfluencies occur. We came up with the idea that there is a very low probability of having

a partial word type of disfluency if the actual word has less than 7 unicode characters.

Among those words which have more than 7 characters, one of them is chosen at random. For that chosen word, a random position is chosen till which the partial word would be created. Thus the partial word formed is our reparandum and is placed before the original word.

- **Phrase Repair**:

  This encircles the typical case of repair disfluency, also called correction. First, we randomly choose 2-6 contiguous words. Then we apply the idea that in a typical correction, the lexical item/POS tag or some related feature of either the first or last word remains the same in reparandum and alteration. Thus keeping either the first or last word unchanged, we modify the rest of the words. To achieve this task, we use Muril (Khanuja et al., 2021) and applied fill mask algorithm sequentially - to get as natural sounding text as possible with respect to the new words that are being generated.

- **Code Mix Repair**:

  Code mix disfluencies is the area where not much has been thought into in the field. We tried to replicate the behaviour of code mix disfluency taking the help of LTRC translation engine [4] .

  Unlike phrase repair where we could simply replace the tokens using fill mask, here we cannot simply replace the tokens with their translations. The main reason being the grammar and sentence structure of the languages involved along with other factors of translation. Hence we translate the whole sentence, and then randomly first k words. This acts as the reparandum, alteration being the full sentence. Note that we only dealt with *Indian Language - English* code mixed data.

5.4. False Start

First, we choose two distinct sentences at random to produce false starts. Subsequently, we divide the initial sentence into two parts at random and join the first segment of the split with the second sentence. With a random probability, instead of simply splitting the sentence, we first translate it, then split and follow the same method, to produce a more natural sounding code mixed false start.

Algorithm 1 shows the entire algorithm for synthetically generating the disfluencies.

[4] https://ssmt.iiit.ac.in/translate

6. Experiment and Methodology

For this task, we used XLM-RoBERTa (Conneau et al., 2019) which is a multilingual version of RoBERTa (Liu et al., 2019). Having about 125M parameters, it is trained on on 2.5 TB of filtered CommonCrawl data in 100 languages with the Masked language modeling (MLM) objective.

We fine-tuned the XLM-RoBERTa model for classification task on our training dataset - which comprises synthetic data as well as real-world annotated data. Table 3 shows the hyperparameters used for the fine-tuning task.

Parameter	Value
Optimizer	Adam
LR	3e-5
Wt Decay	0.01
Batch Size	16
Epochs	10

Table 3: Hyperparameters used for the fine-tuning of XLM-RoBERTa for disfluency identification.

We set the *P_disf* in the Algorithm 1 such that the overall disfluency percentage(by words) in the data stays in the range 8-10% so as to mimic real world data.

7. Performance & Observation

We fine-tuned the model on the generated synthetic data along with real-world annotated data. An additional collection of annotated data from the real world was used for testing. Languages like Hindi and Tamil gave decent results of F1 scores >35. One reason their ratings exceed those of other languages could be attributed to a higher availability of data. Marathi and Telugu gave the lowest of scores, primarily because of less amount of real world data available for them.

Another experiment was run, this time by changing the synthetic augmentation algorithm to produce a better distribution of all the disfluency cate-

Language	Test data (in hrs)	Exp. 1	Exp. 2
Hindi	2	59	60
Bengali	2	22	25
Marathi	1	8	9
Telugu	1	9	9
Kannada	1	16	20
Tamil	2	35	43

Table 4: Weighted F1 scores for the task of disfluency identification.

Algorithm 1 Synthetically augmenting disfluencies

Require:
Sentence or a text on which disfluency needs to be added.
list of filler words *FW*, pet phrases *PP* and edit terms *ET* in each language.
Ensure: text filled with disfluency.
P_disf = A probability which decides whether disfluency will be injected in current sentence or not
if P_disf == True **then**
    D_{type} = randomly choose the type of disfluency to inject
    if D_{type} in (filler words, pet phrase) **then**
        P_{fw} = probability to inject multiple filler words
        P_{pp} = probability to inject multiple pet phrase
        pos = random position to inject filler word/pet phrase
        generate_disfluency($text$, pos, P_{fw}, P_{pp}, FW, PP) ▷ This adds the disfluent words at the position and returns the final synthesized text
    else if D_{type} in (repeat, repair, false start) **then**
        $start\_pos$, end_pos = randomly choose the starting position(word) and the end position.
        $rep\_substring$ = text[$start\_pos$: end_pos] ▷ this substring acts as the alteration
        add_edit_terms($text$, end_pos, FW, PP, ET) ▷ This internally adds edit terms/filler words/pet phrases or a combination of them to the end of the chosen substring
        generate_disfluency($text$, $start\_pos$, end_pos, $rep_substring$) ▷ This adds the generated reparandum before the chosen substring and returns the details of reparandum, alteration and the final synthesized text
    end if
end if

gories. The *P_disf* in the Algorithm 1 was tweaked such that the overall disfluency percentage was around 21%. Additionally, improving the distributions among the various disfluency categories was an important point that we worked on. To get better distribution and more quantity of repairs, the whole list of categories and subcategories of disfluencies were flattened into one list. This ensured that each kind of disfluency (sub)category will have equal probability. The Table 4 shows the weighted F1 scores calculated for both the experiments.

8. Conclusion and Future Work

We have presented detailed guidelines for annotating disfluencies in real-world conversations, accompanied by an algorithm for synthesizing such disfluencies in the data. The necessity of this thorough annotation is underscored by the complexity of the task, as evidenced by the obtained scores, indicating that identifying disfluencies for Indian Languages in continuous real-world conversations poses a significant challenge. Furthermore, the synthetic augmentation process requires constant refinement to better emulate real-world disfluencies. The guidelines outlined in Section 3.3 highlight numerous subtle nuances that models must learn to accurately identify or not identify them as disfluencies.

In our approach, we started with the acquired transcripts and progressed from there. Therefore, it stands to reason that a higher-quality ASR system with as low Word Error Rate (WER) as possible would enhance the efficacy of the entire workflow.

Going forward, there are several directions in this field that has to be explored. It is imperative to continuously refine the annotation criteria in addition to gathering additional datasets, especially from real-world sources. Furthermore, it is expected that some intelligent usage of semantic knowledge pertaining to punctuations, grammatical chunks or part-of-speech tags will improve the algorithm's overall performance. These might have a crucial role both while artificially synthesizing disfluent data as well as for disfluency identification task.

References

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023a. Adversarial training for low-resource disfluency correction. *arXiv preprint arXiv:2306.06384*.

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023b. Disco: A large scale human annotated corpus for disfluency correction in indo-european languages. *arXiv preprint arXiv:2310.16749*.

Marcus Colman and Patrick Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. 2021. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. *arXiv preprint arXiv:2106.04016*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Peter A Heeman and James Allen. 1999. Speech repains, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–572.

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

Rohit Kundu, Preethi Jyothi, and Pushpak Bhattacharyya. 2022. Zero-shot disfluency detection for indian languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4442–4454.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. Lard: Large-scale artificial disfluency generation. *arXiv preprint arXiv:2201.05041*.

Sharath Rao, Ian Lane, and Tanja Schultz. 2007. Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. In *Proceedings of Machine Translation Summit XI: Papers*.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Shaolei Wang, Wangxiang Che, Qi Liu, Pengda Qin, Ting Liu, and William Yang Wang. 2020. Multi-task self-supervised learning for disfluency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9193–9200.

Wen Wang, Gokhan Tur, Jing Zheng, and Necip Fazil Ayan. 2010. Automatic disfluency removal for improving spoken language translation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5214–5217. IEEE.