Disfluency Detection in English-Hindi Languages using Deep Models

4th-Year Final Project (Initial)

Name: Amit Mondal

Roll: 302110501002

Class: BCSE-IV

Supervisors: Professor Sivaji Bandyopadhyay,

Jadavpur University, Kolkata

Acknowledgement

I would also like to express my gratitude to **Professor Sivaji Bandyopadhyay** for their unwavering support and encouragement. Their commitment to excellence and passion for advancing research in disfluency detection have been a driving force behind the success of this project. Their dedication to fostering a collaborative and intellectually stimulating academic environment has created an atmosphere conducive to innovation and academic growth.

Furthermore, I extend my appreciation to **Professor Dipankar Das** for generously sharing their expertise in Natural Language Processing, Social Networks, Sentiment / Emotion Analysis, Information Extraction. Their profound knowledge and willingness to engage in meaningful discussions have significantly enriched the depth and breadth of this research.

In acknowledging **Professor Sivaji Bandyopadhyay** 's pivotal role in this project, I am reminded of the importance of mentorship in academic pursuits. Their guidance has been instrumental in shaping my research skills and scholarly approach. I am sincerely grateful for the opportunity to work under the mentorship and for the lasting impact they have had on both my academic and professional development.

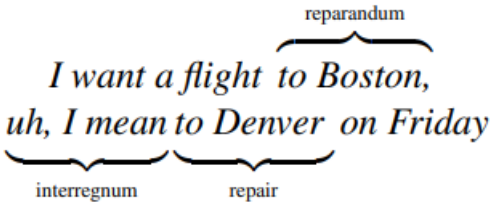
Abstract

Disfluency correction is the process of removing disfluent elements like fillers, repetitions and corrections from spoken utterances to create readable and interpretable text. DC is a vital post-processing step applied to Automatic Speech Recognition (ASR) outputs, before subsequent processing by downstream language understanding tasks. Existing research has primarily focused on English due to the unavailability of large-scale open-source datasets. Towards the goal of multilingual disfluency correction, we present a high-quality human-annotated corpus covering four important Indo-European languages: English, Hindi. We provide extensive analysis of results of state-of-the-art the models across all two languages obtaining F1 scores of 91.19 (English), 82.79 (Hindi).

Contents

1.Introduction

Disfluency informally refers to any interruptions in the normal flow of speech, including false starts, corrections, repetitions and filled pauses. Shriberg defines three distinct parts of a speech disfluency, referred to as the reparandum, interregnum and repair. As illustrated in Example , the reparandum to Boston is the part of the utterance that is replaced, the interregnum uh, I mean (which consists of a filled pause uh and a discouse marker I mean) is an optional part of a disfluent structure, and the repair to Denver replaces the reparandum. The fluent version is obtained by removing reparandum and interregnum words although disfluency detection models mainly deal with identifying and removing reparanda. The reason is that filled pauses and discourse markers belong to a closed set of words and phrases and are trivial to detect.



Example: three distinct parts of a speech disfluency

Disfluency Type	Description	Examples
Filler	Words like uhh, err, uhmm that are often uttered to retain turn of speaking. Each language has a different set of filler words commonly uttered	EN: Write a message to um Sarah. HI: ईमेल डॉ. जॉनसन उम्म हाँ बीसीसी डेव।
Repetition	Consists of words or phrases that are repeated in conversational speech	EN: Add this number to my to my contacts. HI: पापा को जल्दी से जल्दी से टेक्स्ट भेजो।
Correction	Disfluencies that consist of words incorrectly spoken and immediately corrected with a fluent phrase	EN: Get me the order my order status on the desk chair I ordered from Overstock. HI: तनु को एक उम्म एक टेक्स्ट मैसेज भेजो।
False Start	Examples where the speaker changes their chain-of-thought mid sentence to utter a completely different fluent phrase	EN: In an email let's email Tom Hardy about Saturday's video shoot. HI: रोहम अरे रोहन को ईमेल में इमेज अटैच्ड करो सेंड करो।
Fluent	Examples which do not contain any disfluent words or phrases	EN: Can you make a note for Johnny that says dinner at eight on my laptop? HI: क्या राकेश को ई-मेल भेज सकते हो I

Table-1 : Types of sentences observed in the corpus. All disfluencies are marked in red; EN-English, HI-Hindi.

2.Related Work

Approaches to disfluency detection task fall into three main categories: noisy channel models, parsing-based approaches and sequence tagging approaches. Noisy channel models use complex tree adjoining grammar (TAG) based channel models to find the “rough copy” dependencies between words. The channel model uses the similarity between the reparandum and the repair to allocate higher probabilities to exact copy reparandum words. Using the probabilities of TAG channel model and a bigram language model (LM) derived from training data, the noisy channel models generates n-best disfluency analyses for each sentence at test time. The analyses are then reranked using a language model which is sensitive to the global properties of the sentence, such as a syntactic parser based LM.

Parsing-based approaches detect disfluencies while simultaneously identifying the syntactic structure of the sentence. Typically, this is achieved by augmenting a transition-based dependency parser with a new action to detect and remove the disfluent parts of the sentence and their dependencies from the stack. Joint parsing and disfluency detection can compare favorably to pipelined approaches, but requires large annotated treebanks containing both disfluent and syntatic structures for training.

My proposed approach, based on an multilingual disfluency correction belongs to the class of sequence tagging approaches. These approaches use classification techniques such as conditional random fields, hidden Markov models and deep learning based models to label individual words as fluent or disfluent. Sequence tagging methods work well for disfluency removal from real-life spoken utterances, assigning disfluent/fluent label to every word in the sentence.

3.Dataset Details

This section analyzes the DISCO corpus, This is created with the help of English, Hindi, German and French language experts. DISCO contains parallel disfluent-fluent sentence pairs in the above four languages and English translations of fluent sentences in Hindi and German along with disfluency and domain labels. My experiments has primarily focused on English and hindi data.

3.1 Types of Disfluency

There are four types of disfluencies observed in the dataset: Filler, Repetition, Correction and False Start. Additionally, there are some fluent sentences present in the corpus. Table-1 describes each type of sentence with some real examples from the dataset.

3.2 Data Collection Method

I can extract disfluent sentences and domain labels in English, Hindi from DISCO corpus. These utterances consist of human dialogues like making notes, monitoring fitness, adding new contacts, opening apps, etc. All sentences are shared with respective language experts for fluent sentence creation and disfluency-type annotation.

Lang	No. of sentence pairs	No. of words	% disfluent words
English	3980	39690	15%
Hindi	3171	32299	19%

Table: Total count of disfluent-fluent pairs in DISCO and percentage of disfluency present;

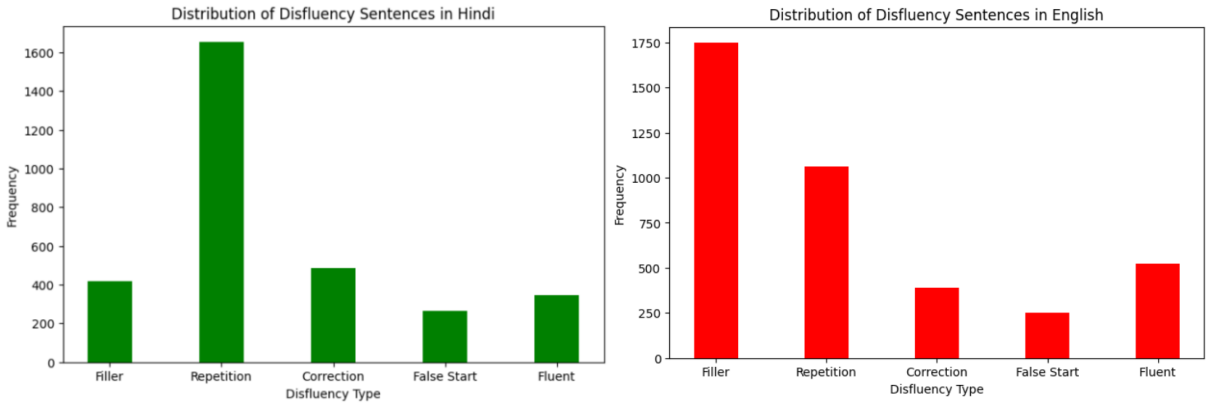


Figure: Distribution of sentences across disfluency types for two languages in DISCO.

Lang	Mean length of disfluent sentences	Mean length of fluent sentences
English	11.24 words	9.61 words
Hindi	10.38 words	8.34 words

Table: Average length of disfluent and fluent utterances in the DISCO corpus for each language.

4. Experiments

4.1 Data Processing

All parallel sentence pairs are passed through a punctuation removal module to reduce the number of tokens for classification. As per the structure of disfluencies, consider fluent terms to always follow disfluent terms in an utterance. Disfluent utterances are marked with the positive label (1) and fluent utterances with the neutral label (0).

4.2 Baseline Models

I am use a combination of smaller ML models, larger transformer models. All models are trained on an 80:10:10 train: valid: test split for each language.

4.2.1 ML Baselines

I am experiment with using Conditional Random Fields (CRFs) and Recurrent Neural Network (RNN) based techniques for token classification in Disfluency Correction. These models require fewer labeled data and are ideal for low-resource domain-specific training. Token-level features from a powerful multilingual transformer, XLM-R ,were used for finetuning the CRF and RNN models.

4.2.2 Transformer Baselines

Transformers are large and powerful neural networks capable of learning complex text representations for many downstream NLP tasks. I am experiment with one multilingual transformers: XLM-R. Finetuning for sequence tagging is performed by adding a classification head (on top of these transformers) that performs sub-word level binary prediction. Prediction of a word to be disfluent/fluent is the prediction of the first sub-word to be disfluent/fluent.

Multilingual Named Entity Recognition (NER) is a challenging task in natural language processing that requires models capable of understanding and classifying entities across various languages. I focus on implementing a Transformer baseline for Multilingual NER using the SimpleTransformer model. The goal is to enhance adaptability and performance in recognizing named entities in diverse linguistic contexts.

4.3 Experimental Setup

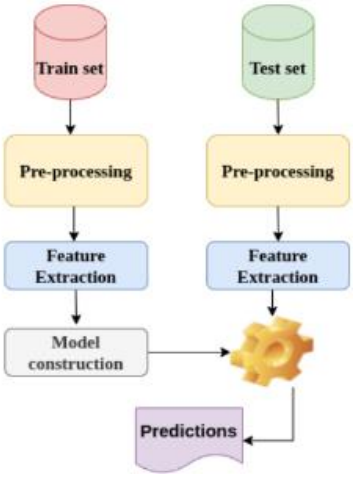
CRF and RNN models are trained using the Flair-NLP framework till the validation cross-entropy loss saturates. We start with a learning rate of 0.1 and reduce it by half each time the model does not improve for three consecutive epochs. Transformer models are trained using the popular transformers package. We use a learning rate of 2e-5 and a weight decay of 0.01. XML-R transformer models are trained for 40 epochs and SimpleTransfomer (NER) model are trained for 1 epoch and 5 epochs using the Adam optimizer.

4.3.1 Hardware support

All ML and Transformers baselines were trained with T4 GPUs provided by Google Colab.

4.4 Methodology

The disfluency detection task is modeled as a sequence labeling task using a CRF model trained with text based features. The methodology involves Pre-processing, Feature Extraction and Model Construction. Framework of the proposed methodology is shown in Figure 1 and the steps involved are explained below:



5. Results and Analysis

All results are reported using the F1 score metric. Combined results across languages are described in table. As the complexity of models increases, the overall accuracy also increases. As the complexity of models increases, the overall accuracy also increases. Transformer architectures perform better than CRF and RNN-based models consistently. In each language, the best models produce 90+ F1 scores on blind test sets, indicating that our corpus successfully solves the data scarcity problem.

Model	English	Hindi
CRF	86.91	73.89
RNN	87.62	79.11
NER	91.19	82.59
XML-R	99.0	98.0

Performance across disfluency types is described in table. We observe that the model performs poorly for fluent sentences in English due to fewer samples in the test set. In Hindi, false starts are the most difficult disfluencies to correct. Further examination reveals that our model often under-corrects longer false starts, especially in the presence of other disfluencies like fillers. Our model performs robustly across all domain types of utterances. Readers are strongly urged for domain-level analysis of Disfluency Detection results. Although existing datasets are of diverse domains, our experiments show that models trained on DISCO outperform test sets from other DC datasets.

```
sent1 = "Add sandwich Add sandwich for lunch"
sent2 = "Add 700 calories a 700 calorie protein shake for breakfast"
sent3 = "take note take note of my vitamin consumption"
sent4="Log intake log calorie intake"
sent5="Today Todays is Sunday "
```

```
[
[{'Add': '1'}, {'sandwich': '1'},{'Add': '0'},{'sandwich': '0'},{'for': '0'}, {'lunch': '0'}],
[{'Add': '0'}, {'700': '1'},{'calories': '0'},{'a': '0'},{'700': '0'},{'calorie': '0'},{'protein': '0'},{'shake': '0'},{'for':
'0'},{'breakfast': '0'}],
[{'take': '1'},{'note': '1'},{'take': '0'},{'note': '0'},{'of': '0'},{'my': '0'},{'vitamin': '0'},{'consumption': '0'}],
[{'Log': '0'},{'intake': '1'}, {'log': '0'},{'calorie': '0'}, {'intake': '0'}],
 [{'Today': '1'}, {'Todays': '0'}, {'is': '0'}, {'sunday': '0'}]
]
```

6. Conclusion and Future Work

This experiment introduces the DISCO dataset for disfluency detection in two widely spoken languages: English and Hindi. My work highlights the importance of large-scale projects in NLP that scale the amount of labeled data available. Spoken interactions between humans and AI agents are riddled with disfluencies. Eliminating disfluencies not only improves readability of utterances but also leads to better downstream translations. DISCO dataset, which consists of roughly 3000 parallel disfluent-fluent sentences in each language, significantly reduces the data scarcity problem in DC.

Future work lies in experimenting with better ML models for sequence tagging-based DC supporting multilingual training. These should also incorporate linguistic features like reparandum, interregnum and repair. Multimodal DC presents a promising direction as it has the capability of using both speech and text features for correction tasks. Exploring semi-supervised and unsupervised learning approaches could also alleviate the dependency on extensive labeled datasets, making the system more adaptable and cost-effective. Moreover, addressing ethical considerations, such as privacy and bias concerns, is imperative in ensuring responsible deployment. Lastly, evaluating the real-world applicability of these advancements and their impact on diverse communication scenarios should guide the future trajectory of disfluency detection research.

7. Limitations

My work consists of two limitations. Firstly, since our annotation process consisted of one annotator for each language, we could not report metrics such as inter-annotator agreement. However, since it is a relatively more straightforward task and consists of only removing disfluent words from spoken utterances, the structure of disfluencies helps us recognize disfluency types easily. Secondly, we do not compare trained models on DISCO with other datasets due to varied domain of existing datasets. We found that existing datasets like Switchboard, LARD consisted of utterances from very diverse data sources.

8. References

- Paria Jamshid Lou, Peter Anderson, Mark Johnson, 2018, Disfluency detection using auto-correlational neural networks
- Vicky Zayats, Mari Ostendorf, Hannaneh Hajishirzi, 2016, Disfluency detection using a bidirectional LSTM
- Nikhil Saini, Preethi Jyothi and Pushpak Bhattacharyya, 2021, Survey: Exploring Disfluencies for Speech To Text Machine Translation.
- Vineet Bhat, Preethi Jyothi, Pushpak Bhattacharyya, 2023, DISCO: A Large Scale Human Annotated Corpus for Disfluency Correction in Indo-European Languages.
- Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, Daniel J. Liebling, 2021, Disfluency Detection with Unlabeled Data and Small BERT Models.
- Ganesh Babu C, Sushvin M, 2023, Disfluency Identification for 6 Indian Languages at ICON 23