Abstract

Reduplication is a prominent linguistic phenomenon in Bengali, serving various grammatical and semantic functions such as intensification, plurality, and emphasis. Its complex nature, characterized by structural variations and diverse usages, poses significant challenges for computational processing. This research addresses these challenges through a multi-phase approach aimed at enhancing the accuracy of identifying, generating, and sense-tagging reduplicated forms, which are critical for applications such as machine translation and sentiment analysis.

This research introduces a precise rule-based methodology for detecting partially reduplicated Multi-Word Expressions (MWEs) in Bengali texts. The process is executed in two phases to enhance accuracy and effectiveness. In the first phase, a Levenshtein distance based algorithm identifies partially reduplicated forms by assessing word similarity and flagging relevant instances. The second phase refines this output using a noble technique known as *Word Expansion*. Word Expansion technique, significantly improving detection performance, as reflected by metrics of 90.00% Precision, 85.71% Recall, and an F1-Score of 87.80%.

Additionally, this study leverages supervised machine learning to identify reduplicated MWEs in Bengali corpora. Candidate MWEs are extracted using syntactic rules, followed by classification through a Random Forest algorithm. This approach employs diverse features, such as association measures and linguistic attributes, achieving superior performance with 90.90% Precision, 95.24% Recall, and an F1-Score of 93.02%. These findings highlight the potential of both rule-based and machine learning methodologies in addressing the structural diversity and complexity of reduplication in Bengali texts.

Further, the study models the generation of complete and partial reduplications using two-way finite-state transducers (FSTs). While complete reduplications are effectively modeled with a single FST, partial reduplications require multiple FSTs to accommodate their varied patterns. This process achieves an F1-Score of 88.11%, demonstrating its utility in identifying reduplication instances.

Finally, a robust sense-tagging framework for reduplicated forms leverages BERT-based word embeddings. By clustering word vectors based on cosine similarity scores, the system achieves distinct sense identification, with F1-Scores of 84.96%. These findings underscore the potential of advanced NLP techniques to effectively process and understand the complexities of Bengali reduplicated forms.

This research provides valuable insights into computational linguistics, offering scalable solutions for NLP tasks in low-resource language scenarios.