Malayalam Parser for Dataset Creation

Abstract Presentation

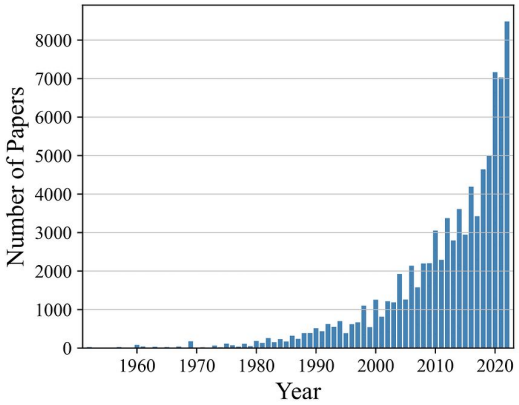
Guided by:
Dr. Mary Priya Sebastian

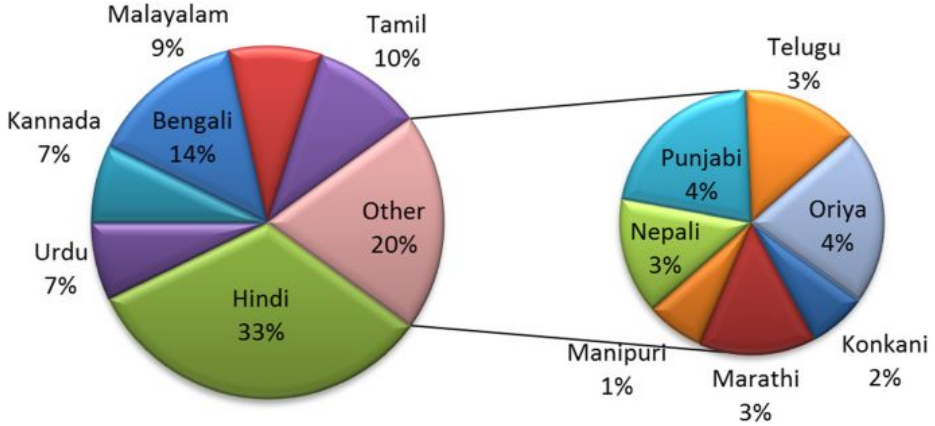
Fathima Jennath N K
Gautham C Sudheer
Godwin Gino
Mohammed Basil

Contents

Natural Language Processing

- **Multimodality**
  - Visual Data in NLP
  - Structured Data in NLP
  - Speech & Audio in NLP
  - Programming Languages in NLP
- **Natural Language Interfaces**
  - Dialogue Systems & Conversational Agents
  - Question Answering
- **Semantic Text Processing**
  - Discourse & Pragmatics
  - Knowledge Representation
  - Text Complexity
  - Semantic Search
  - Representation Learning
  - Word Sense Disambiguation
  - Semantic Parsing
  - Language Models
  - Semantic Similarity
- **Sentiment Analysis**
  - Opinion Mining
  - Stylistic Analysis
  - Intent Recognition
  - Emotion Analysis
  - Aspect-based Sentiment Analysis
  - Polarity Analysis
- **Syntactic Text Processing**
  - Tagging
  - Morphology
  - Chunking
  - Phonology
  - Text Error Correction
  - Text Segmentation
  - Typology
  - Syntactic Parsing
  - Phonetics
  - Text Normalization
- **Linguistics & Cognitive NLP**
  - Linguistic Theories
  - Cognitive Modeling
  - Psycholinguistics
- **Responsible & Trustworthy NLP**
  - Ethical NLP
  - Low-Resource NLP
  - Robustness in NLP
  - Green & Sustainable NLP
  - Explainability & Interpretability in NLP
- **Reasoning**
  - Textual Inference
  - Commonsense Reasoning
  - Numerical Reasoning
  - Knowledge Graph Reasoning
  - Machine Reading Comprehension
  - Fact & Claim Verification
  - Argument Mining
- **Multilinguality**
  - Cross-Lingual Transfer
  - Machine Translation
  - Code-Switching
  - Typology
- **Information Retrieval**
  - Indexing
  - Document Retrieval
  - Semantic Search
  - Text Classification
  - Passage Retrieval
- **Information Extraction & Text Mining**
  - Coreference Resolution
  - Text Clustering
  - Named Entity Recognition
  - Event Extraction
  - Open Information Extraction
  - Term Extraction
  - Text Classification
  - Topic Modeling
  - Summarization
  - Relation Extraction
- **Text Generation**
  - Data-to-Text Generation
  - Summarization
  - Question Generation
  - Dialogue Response Generation
  - Captioning
  - Paraphrasing
  - Machine Translation
  - Speech Recognition
  - Text Style Transfer
  - Code Generation





Introduction

- Importance of Natural Language Processing (NLP) in regional languages
- Focus on the specific relevance of Malayalam in the context of NLP applications.
- Challenges associated with the scarcity of annotated datasets.
- Analyze both the syntactic and semantic structures of Malayalam sentences
- Applications such as sentiment analysis, named entity recognition, etc.
- Potential impact on advancing research and applications specific to the Malayalam

Description of Project

To create a Malayalam Parser for dataset creation, involving data collection, preprocessing, manual annotation, and training using various parsing approaches to address the scarcity of annotated datasets in Malayalam for NLP applications.

Scope of the Project

- Address the scarcity of annotated datasets in the Malayalam language for Natural Language Processing (NLP) applications.
- Analysis of grammatical structures in Malayalam text data.
- Contributing to the overall improvement of Malayalam language processing technologies.
- Does not include specific application development for sentiment analysis, named entity recognition, or machine translation.

Functional Requirements

- Parse and analyze Malayalam language text to identify linguistic components such as words, phrases, and sentences.
- Determine grammatical structure, syntax, and semantics of Malayalam sentences to facilitate accurate linguistic analysis.
- Provide functionality for part-of-speech tagging, syntactic parsing, and semantic analysis tailored for the Malayalam language.

Functional Requirements

- Support for handling compound words, inflections, and variations in word forms commonly found in Malayalam text.
- Generation of a part-of-speech tagged dataset, named entity dataset, and sentimental tagged dataset, contributing to the advancement of language processing technologies in Malayalam
- Implement a user-friendly interface that allows users to input Malayalam text for analysis

System Features

- Text Parsing: The parser breaks down Malayalam language text into its constituent components such as words, phrases, and sentences.
- Lemmatization and Morphological analysis: Identifying the base forms of words and analyzing word forms to determine grammatical properties in Malayalam text .
- Dependency Parsing: Identifying the syntactic dependencies between words in a Malayalam sentence.

System Features

- Semantic Analysis: The system determines the meaning and interpretation of Malayalam text, capturing semantic relationships between words and phrases.
- Part-of-Speech Tagging: Each word in a Malayalam sentence is assigned appropriate part-of-speech tags, indicating its grammatical function.
- Syntactic Parsing: It analyzes the syntactic structure of Malayalam sentences, identifying constituents and their hierarchical relationships.
- Error Handling: Mechanisms are in place to detect and handle errors or inconsistencies in input Malayalam text, ensuring parsing reliability.

Software / Hardware Requirements

- Windows 10 or later
- MacOS 10.13 High Sierra or later
- Ubuntu 18.04 LTS or later
- A modern processor (e.g., Intel Core i5 or equivalent)
- Sufficient RAM (at least 4GB)
- Available storage space for software installation
- Python (version 3.6 or later)
- Other programming languages and frameworks suitable for NLP development like NLTK, spaCy, scikit-learn, TensorFlow, etc. may be necessary

Conclusion

A comprehensive Malayalam language processing tool facilitating accurate linguistic analysis and dataset generation for NLP applications.

- Parsing and analysis of Malayalam text, enabling identification of linguistic components and determination of grammatical structure, syntax, and semantics
- Generates part-of-speech tagged, named entity, and sentiment-tagged datasets
- Contribute significantly to the advancement of language processing technologies in Malayalam.

Thank you