Malayalam Parser for Dataset Creation

Guided by:
Dr. Mary Priya Sebastian

Fathima Jennath
Gautham C Sudheer
Godwin Gino
Mohammed Basil

Contents

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

Problem Definition

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.

Objectives

- Data Collection
  a) Gather text data from diverse sources in Malayalam language
  b) Aim for a sufficient volume of data to represent the language's usage patterns adequately
- Data Preprocessing
  a) Perform tokenization, normalization, and cleaning of the collected data
  b) Handle any inconsistencies or noise in the data to ensure quality

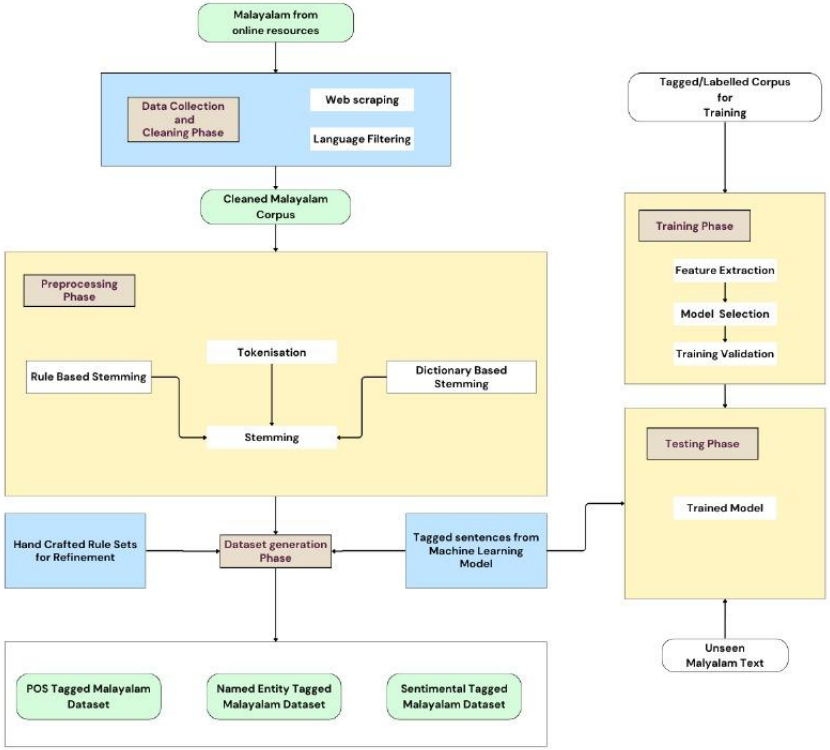
Objectives

- Manual Annotation
  a. Annotate a representative subset of the preprocessed data with grammatical and syntactic information
  b. Employ linguistic experts or proficient annotators to ensure accurate annotations.
- Parser Development
  a. Train the parser using the annotated dataset to understand Malayalam syntax and semantics

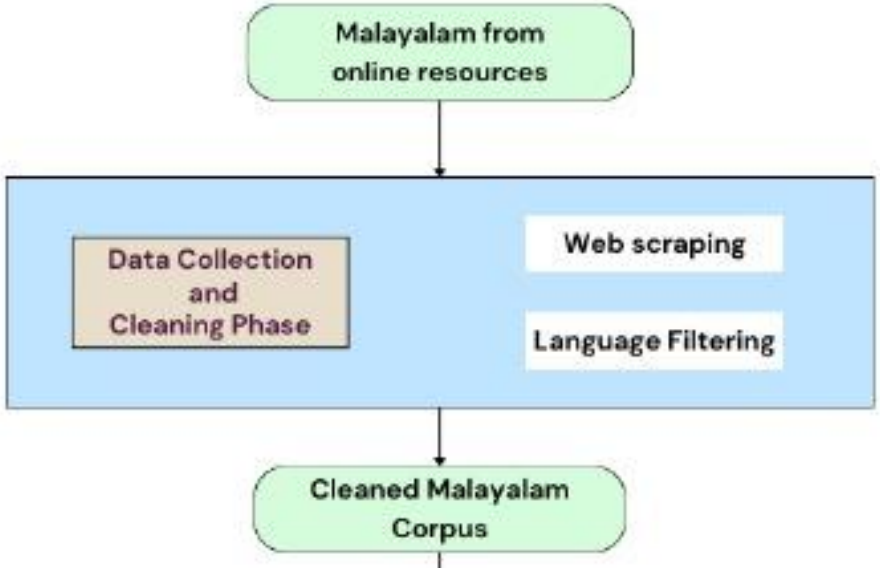
Scope and Relevance

- 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.

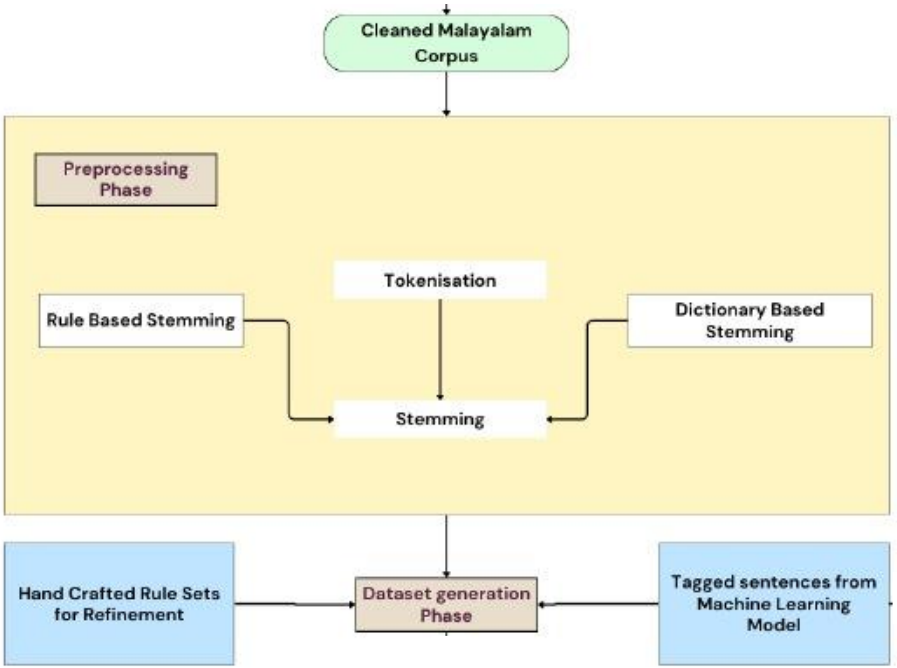
System Design



Data Collection and Cleaning Phase



Data Preprocessing



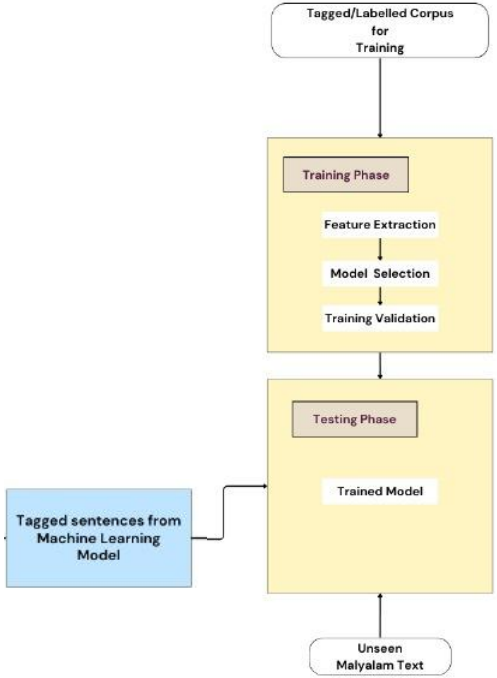
Prepositions- (?:\b(?:ഉപസർഗങ്ങൾ)\b)

Conjunctions- \b(?:ഉം|അല്ലെങ്കിൽ|അഥവാ|അല്ല|അതിനും|അങ്ങനെ)\b

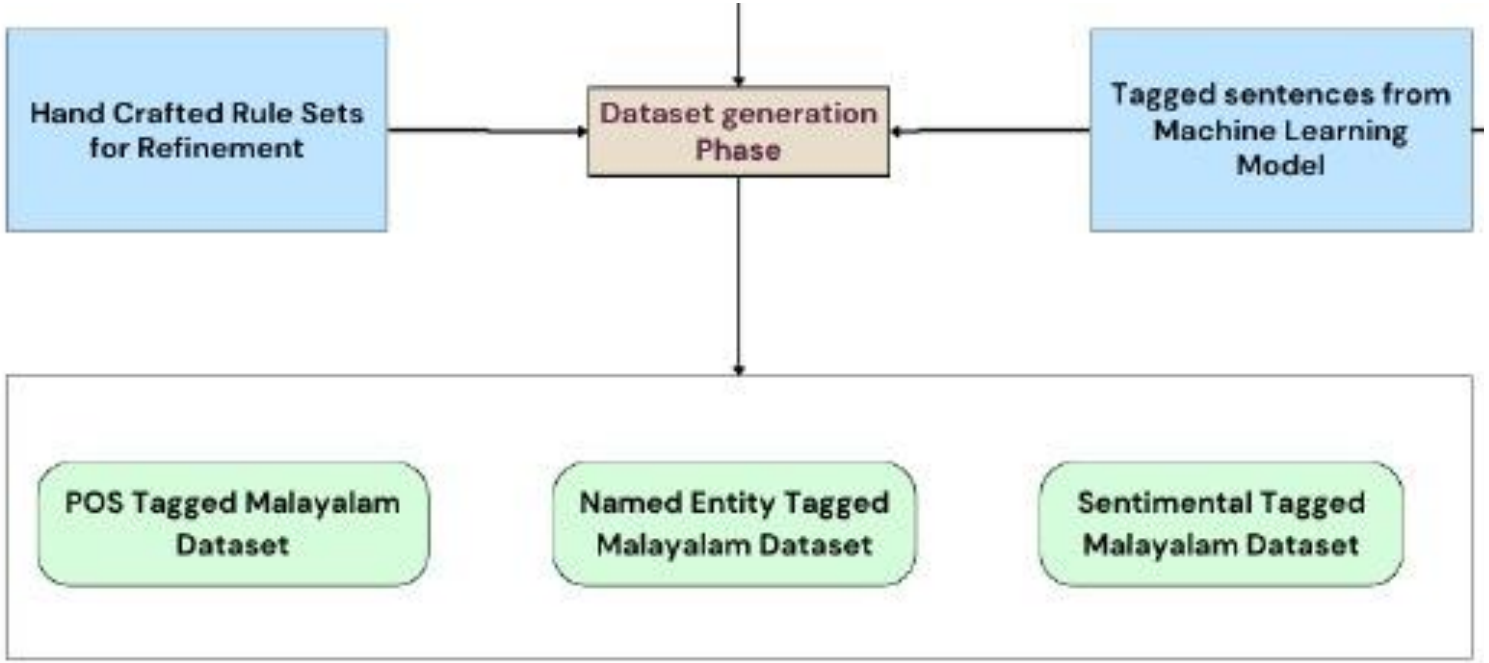
Determiners - \b(?:
ഈ|അത്|അതിന്റെ|ഇവയ്ക്ക്|അവയ്ക്ക്|ആ|അവന്റെ)\b

Pronouns-\b(?:
ഞാൻ|നീ|അവൻ|അവൾ|അത്|ഞങ്ങൾ|അവർ|എന്റെ|നിന്റെ|അവർക്ക്|
എന്റെ|അവർക്ക്|ആർ|ഏത്|എങ്ങനെ|ഏതാണ്)\b

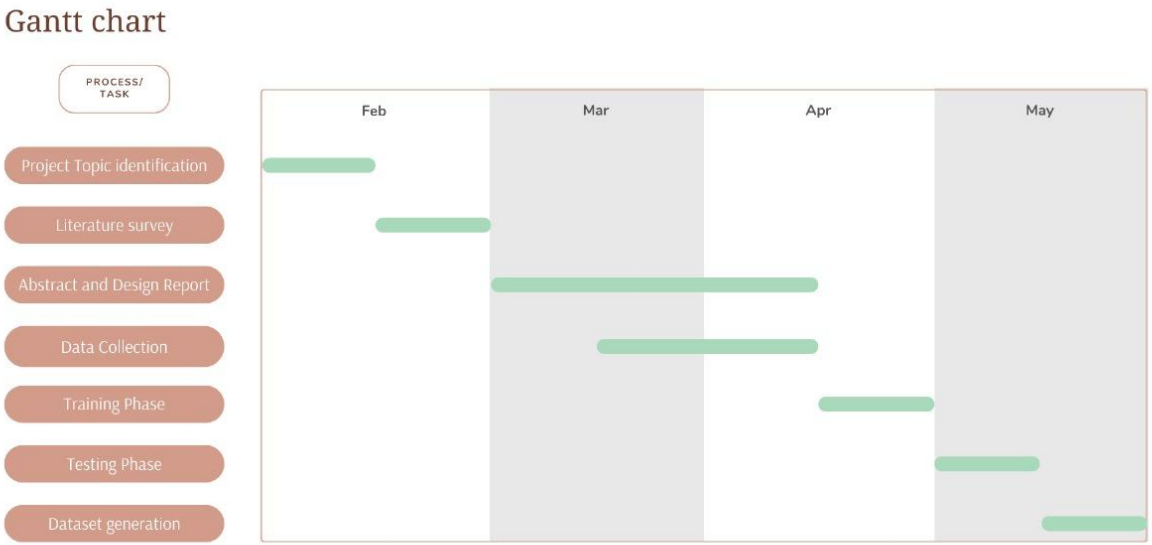
Training and Testing



Output


Hand Crafted Rule Sets for Refinement
Dataset generation Phase
Tagged sentences from Machine Learning Model
POS Tagged Malayalam Dataset
Named Entity Tagged Malayalam Dataset
Sentimental Tagged Malayalam Dataset

Work Division



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

Results

Malayalam Parser

Step into a world of linguistic exploration! Dive into our webpage to uncover existing datasets and transform your input into a mosaic of named entities, POS tags, and sentiment analysis.

അന്ധർക്ക് അനായാസമായി വായിക്കാൻ പ്രാപ്തി നൽകുന്ന റിയ വികസിപ്പിച്ച ഉൽപ്പന്നത്തിന് ഗൂഗിൾ സമ്മാനം നൽകി

Submit

Entity	Tag
ഗൂഗിൾ	Organization
റിയ	Person

Token	POS Tag
ഗൂഗിൾ	Proper Noun
സമ്മാനിച്ചു	Verb
ഉൽപ്പന്നം	Noun
വികസിപ്പിച്ചെടുത്തു	Verb
റിയ	Proper Noun
എന്ന്	Pronoun
അന്ധൻ	Adjective
വായിച്ചു	Verb
എളുപ്പത്തിൽ	Adverb

Sentimental Analysis
Positive

Click here to download and use our data sets

Named Entity ⬇ POS Tagged ⬇ Sentimental ⬇

Results

Malayalam Parser

Step into a world of linguistic exploration! Dive into our webpage to uncover existing datasets and transform your input into a mosaic of named entities, POS tags, and sentiment analysis.

Mary was awarded the best student at Rajagiri College.

Submit

Entity	Tag
Mary	Person
Rajagiri College	Organization

Token	POS Tag
Mary	Proper Noun
was	Auxiliary
awarded	Verb
the	Determiner
best	Adjective
student	Noun
at	Adposition
Rajagiri	Proper Noun
College	Proper Noun
.	Punctuation

Sentimental Analysis
Positive

Click here to download and use our data sets

Named Entity ⬇ POS Tagged ⬇ Sentimental ⬇

Results

Malayalam Parser

Explore More

NOUN

A noun is a word that names a person, place, thing, or idea. It's like a label we use for everything around us. For example, "dog," "cat," "house," and "love" are all nouns. Nouns can be common, like "book" or "table," which are general things, or they can be proper, like "Mary" or "London," which are specific names. In a sentence, nouns can be the subject (the thing doing the action) or the object (the thing receiving the action). They help us talk about the world and communicate with others.

നാമം

നാമം ഒരു വ്യക്തി, സ്ഥലം, വസ്തു, അല്ലെങ്കിൽ ധാരണയെ സൂചിപ്പിക്കുന്നു. നാമങ്ങൾ , സാധാരണ വസ്തുക്കളുടെ പേരാണ് (ഉദാഹരണത്തിന്, പുസ്തകം, ടേബിൾ), അല്ലെങ്കിൽ സ്പഷ്ടമായ, പേരുകളുടെ പേരാണ് (ഉദാഹരണത്തിന്, മേരി, ലണ്ടൻ).ഒരു വാക്യത്തിൽ, നാമങ്ങൾ വാക്യത്തിന്റെ പ്രധാനം അല്ലെങ്കിൽ വാക്യം ചെയ്യുന്ന കാര്യം വിവരിക്കുന്നു. അവ വിഷയം (ചെയ്യുന്ന കാര്യം) അല്ലെങ്കിൽ ഉദ്ദേശം (വിശ്വസിക്കുന്ന കാര്യം) എന്നിങ്ങനെ ഉപയോഗിക്കാം. നാമങ്ങൾ നമ്മുടെ ആശയങ്ങളെ അടിസ്ഥാനമാക്കുകയും, പറയുന്നതിനു സഹായകമാക്കുകയും ചെയ്യുന്നു.

ഉദാഹരണം

- ആകാശം
- വീട്
- മേരി
- ജോൺ
- പച്ച
- ഹോട്ടൽ

Results

Malayalam Parser

- Adjective
- Adposition
- Adverb
- Auxiliary
- Coordinating Conjunction
- Determiner
- Interjection
- Noun
- Numeral
- Other
- Particle
- Pronoun
- Proper Noun
- Punctuation
- Subordinating Conjunction
- Symbol
- Verb

Results

- Implemented basic functionality for processing Malayalam text, including extracting entities, part-of-speech (POS) tags, and sentiment analysis
- Integration with external services such as Google Translator and TextBlob has enabled language translation and sentiment analysis capabilities
- Detect whether the input text is in Malayalam or English, enabling appropriate processing based on the language
- Processed data, including entities, POS tags, and sentiment analysis results, are stored in CSV files for further analysis and reference.

Results

- A basic user interface (UI) is provided through a Django web application, allowing users to input text and receive processed results
- Includes educational resources that provide descriptions and examples for various parts-of-speech (POS) tags in the Malayalam language. These pages serve as valuable reference materials for users interested in understanding the linguistic nuances of Malayalam text

Future Enhancements

- Explore and implement more advanced parsing techniques, such as dependency parsing or deep learning, to enhance the accuracy and robustness
- Implementing a rule-based parsing approach that involves defining grammatical rules and patterns specific to the Malayalam language to improve parsing accuracy and coverage
- Exploring a hybrid approach that combines rule-based and ML techniques to leverage the strengths of both methodologies. For example, using rule-based parsing for deterministic tasks and ML models for probabilistic tasks to achieve a balance between accuracy and flexibility.

Future Enhancements

- Implement error correction mechanisms to handle inaccuracies in the parsing results and provide users with feedback options to report errors and improve the quality of the parser over time
- Incorporate language-specific features and linguistic resources tailored to Malayalam, such as lexicons, morphological analyzers, and syntactic parsers
- Engaging with linguists, researchers, and the local community to gather feedback, validate parsing results, and prioritize future development efforts

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.

References

- Asopa, S., and Sharma, N. (2021) A Hybrid Parser Model for Hindi Language. Indian Journal of Computer Science and Engineering (IJCSE), Vol. 12(1).
- Chen, D., and Manning, C. D. (2014). A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- Nair, L. R. (2013). Language Parsing and Syntax of Malayalam Language. 2nd International Symposium on Computer, Communication, Control and Automation (3CA 2013).
- Berger, A. L., Della Pietra, V. J., and Della Pietra, S. A. (1996). A Maximum Entropy Approach to Natural Language Processing. Association for Computational Linguistics, Vol 22(1).
- Mestry, A., Shende, S., Mahadik, A., and Virnodkar, S. (2014). A Parser: Simple English Sentence Detector and Correction. International Journal of Engineering Research and Technology (IJERT).

References

- Sethi, N., Agrawal, P., Madaan, V., and Singh, S. K. (2016). A Novel Approach to Paraphrase Hindi Sentences using Natural Language Processing. Indian Journal of Science and Technology, Vol 9(28).
- Smith, D. A., and Eisner, J. (2008). Dependency Parsing by Belief Propagation. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Page 145- 156.
- Bharati, A., Kulkarni, A., and Chaudhury, S. (2007). English Parsers: Some Information- based Observations.
- Jayan, J. P., and R, R. (2009). A Morphological Analyzer for Malayalam - A Comparison of Different Approaches. International Journal of Computer Science and Information Technology. Vol 2(2), Page 155-160.
- Vaidya, A., Choi, J. D., Palmer, M., and Narasimhan, B. (2011). Analysis of the Hindi Proposition Bank using Dependency Structure. Proceedings of the Fifth Law Workshop (LAW V), Page 21-29.

References

- Rajan, M., T.S, R., and Bhojane, V. (2014). Information Retrieval in Malayalam Using Natural Language Processing. International Journal of Scientific and Engineering Research, Vol 5(6)
- Rajan, M., Thirumalai, R., and Kumar, V. (2006). Development of a Tamil Parser using Natural Language Processing Techniques. A survey of the state of the art in tamil language technology Vol 6(10).
- Venkatesh, R., Kumar, S., and Arumugam, P. (2014). Building a Lexical Analyzer for Tamil Texts using NLP Approaches. 2014 International Conference on Advances in ICT for Emerging Regions (ICTer).
- Thavareesan, S., and Mahesan, S. (2019). Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation.2019 IEEE 14th Conference on Industrial and Information Systems (ICIIS).

References

- Pai, T. V., Devi, J. A., and Aithal, P. S. (2020). A Systematic Literature Review of Lexical Analyzer Implementation Techniques in Compiler Design. International Journal of Applied Engineering and Management Letters (IJAEML), Vol 4(2), Page 285-301.
- Simmons, R. F., and Burger, J. F. (1968). A Semantic Analyzer for English Sentences. Mechanical Translation and Computational Linguistics, Vol 11.

Thank you