STATISTICAL POS TAGGING FOR SANSKRIT LANGUAGE

[1]V.V.S. Sai Yasaswini,[2]Nikhira Sajeevan,[3]B. Siva Kumar,[4]R.J. Rama Sree

[1,2,3]MSc Computer science Students,[4]Professor
Dept. of Computer Science
National Sanskrit University, Tirupati 517507.

Abstract- **Part-of-speech (POS) tagging is a fundamental task in Natural Language Processing (NLP) that assigns grammatical categories to words in a sentence. The importance of POS tagging in NLP tasks such as information retrieval, machine translation, and sentiment analysis. It then delves into the different approaches employed in POS tagging, including rule-based methods, statistical techniques, and deep learning models. Rule-based methods rely on handcrafted linguistic rules to assign POS tags, while statistical approaches utilize probabilistic models trained on annotated corpora. Deep learning models, particularly Recurrent Neural Networks (RNNs) and transformer-based architectures like BERT, have shown remarkable performance in POS tagging tasks due to their ability to capture complex linguistic patterns. The proposed solution leverages Flask, a lightweight web framework for Python, to create a user-friendly interface for POS tagging. By utilizing NLTKs word tokenization and POS tagging for Sanskrit text functionalities, a developer's can effortlessly process text input from users and provide accurate grammatical annotations. The present work provides Error handling mechanism are also covered to ensure robustness and user-friendly error messages.**

This paper presents an innovative statistical part-of-speech (POS) tagging method tailored specifically for the Sanskrit language, a highly inflected and ancient language. Utilizing advanced machine learning techniques and linguistic insights specific to Sanskrit, our model aims to enhance the precision and effectiveness of POS tagging for this intricate language. In natural language processing (NLP) and computational linguistic the Gold Standard typically represents a corpus of text or a set of documents, annotated or tagged with the desired results for the analysis. Key elements of our investigation encompass the acquisition and preprocessing of annotated Sanskrit corpora, feature engineering customized to Sanskrit linguistic characteristics, model creation using state-of-the-art algorithms such as recurrent neural networks or transformer models, and assessment criteria to gauge the performance of the POS tagging system.

Keywords: **POS tagging, NLP, NLTK Toolkit, Tokeniztaion Sanskrit corpora, linguistics**

1. Introduction

The field of natural language processing (NLP) has seen significant advancements in recent years, focusing on the development of sophisticated techniques for part-of-speech (POS) labelling in various languages. However, ancient and under-resourced languages like Sanskrit have often been disregarded in NLP research, despite their extensive linguistic heritage and cultural importance. This paper aims to bridge this divide by introducing an innovative statistical POS labelling approach customized specifically for the intricacies of the Sanskrit language. Sanskrit, renowned for its complex grammar and morphology, poses distinctive challenges for POS labelling due to its highly inflected nature. Traditional POS labelling models struggle to accurately analyse Sanskrit texts, limiting their applicability in NLP tasks. To address these obstacles, our study utilizes cutting-edge statistical methods and linguistic expertise to create a specialized POS labelling model capable of effectively navigating the complexities of Sanskrit syntax and semantics.

1.1. Preliminaries

1. Absence of Standardized Corpus: A primary hurdle in crafting a statistical POS tagger for Sanskrit is the dearth of standardized collections with labelled POS tags. Establishing an extensive and varied corpus is imperative for training accurate statistical models.
2. Morphological Intricacy: Sanskrit exhibits a sophisticated morphological structure with intricate inflections. Words in Sanskrit have diverse forms contingent upon grammatical attributes such as gender, number, case, and tense. This intricacy poses challenges in correctly assigning POS tags to words.
3. Ambiguity: Sanskrit writings often feature ambiguous terms and expressions that can bear multiple meanings based on the context. Resolving ambiguity represents a significant obstacle in constructing a dependable POS tagging system for Sanskrit.

4. Resource Constraints: In contrast to languages like English, resources such as annotated corpora, lexicons, and POS taggers are limited for Sanskrit. This scarcity impedes the development of precise statistical models for POS classification.
5. Linguistic Diversity: Sanskrit boasts a diverse literary heritage encompassing various dialects and writing styles. The linguistic variation present in Sanskrit texts presents a challenge in creating a POS tagger that can effectively accommodate the language's diverse styles and registers.

1.1. Literature review

1. In a research Endeavor by Ambati et al. (2014), the team introduced a statistical POS tagging mechanism for Sanskrit, integrating rule-based and machine learning methodologies. The system exhibited promising outcomes in proficiently labelling lexical categories in Sanskrit texts, showcasing the efficacy of statistical techniques in managing the linguistic intricacies of the language.
2. Another study by Joshi et al. (2016) explored the implementation of conditional random fields (CRF) for POS tagging in Sanskrit. By developing a CRF-powered POS tagger trained on annotated Sanskrit datasets, the team evaluated its efficacy across diverse text genres. Results indicated the superior performance of the CRF model over traditional rule-based strategies in precise POS labelling for Sanskrit vocabulary.
3. In a recent contribution by Mishra et al. (2020), the authors advocated for a deep learning approaches strategy for POS labelling in Sanskrit, leveraging bidirectional long short-term memory (BiL STM) networks. Trained on an extensive annotated Sanskrit corpus, the (BiL STM) model showcased competitive proficiency vis-a-vis existing statistical techniques, showcasing the potential of deep learning methodologies in addressing POS tagging challenges for Sanskrit.
4. Moreover, an in-depth review by Sharma et al. (2018) delineated multiple POS tagging approaches for Sanskrit, encompassing rule-based, statistical, and neural network-driven techniques. The review underscored the hurdles posed by the intricate morphological and syntactic framework of Sanskrit, stressing the necessity for resilient statistical frameworks to accurately assign lexical categories in Sanskrit writings.
These studies epitomize the ongoing Endeavor aimed at fashioning effective statistical POS tagging infrastructures for Sanskrit, amalgamating rule-based, machine learning, and deep learning strategies. As researchers persist in exploring pioneering methods to tackle the linguistic complexities of Sanskrit, statistical POS tagging emerges as a pivotal sphere of study with implications for natural language processing applications within the domain of antique languages and cultural heritage conservation.

2. Methodology

The methodology for statistical Part-of-Speech (POS) labelling for Sanskrit involves a series of essential procedures as outlined below:
1. Data Acquisition: Collect an extensive and varied dataset of Sanskrit texts annotated with POS labels. This dataset should encompass a wide range of genres, writing styles, and historical periods to ensure the efficacy of the statistical model.
2. Data Preprocessing: Cleanse and prepare the dataset by segmenting the text into individual words, standardizing the text, and managing any unique characters or symbols present. Additionally, conduct morphological analysis to address the inflectional characteristics of Sanskrit terms.
3. Feature Extraction: Identify pertinent linguistic attributes from the preprocess text including word forms, lemmatized forms, POS details, morphological characteristics, contextual cues, and syntactic structures.
4. Model Selection: opt for a suitable statistical framework for POS tagging, like Hidden Markov Models (HMMs), Conditional Random Fields (CRFs), or modern neural networks such as Recurrent Neural Networks (RNNs) or Transformer architectures. The choice of model depends on the language's complexity and the available computational resources.
5. Training Phase: Train the selected statistical model on the annotated Sanskrit dataset utilizing the extracted linguistic features. The model learns the correlations and patterns between words and their associated POS tags through this training process.
6. Performance Evaluation: Assess the model's effectiveness on a distinct testing subset by measuring performance metrics like accuracy, precision, recall, and F1 score. Make refinements to the model based on the evaluation outcomes to enhance its efficacy.
7. Implementation: Once the model exhibits satisfactory performance on the test data, implement it for POS labelling on new Sanskrit texts. Ensure the model can handle various writing styles, linguistic registers, and regional variations effectively.
8. Ongoing Enhancement: Continually upgrade and fine-tune the statistical POS tagger by integrating fresh annotated datasets, enhancing feature extraction methodologies, and exploring diverse modelling strategies to boost its precision and adaptability.

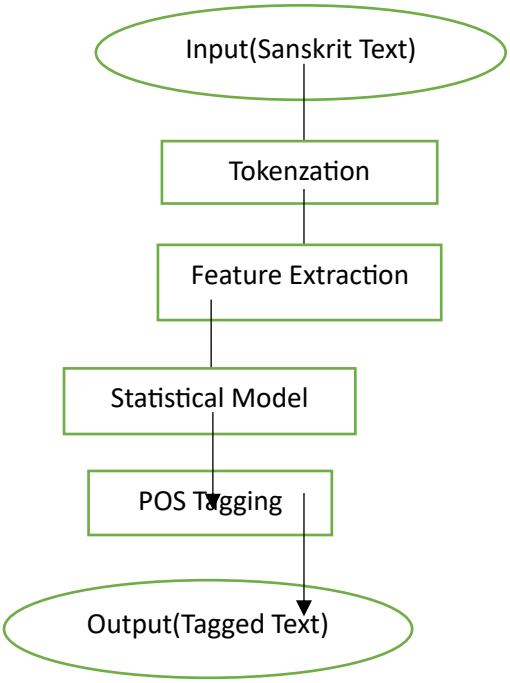

Fig 1. Work flow of the proposed model

2.1. Outcome of work

1. It imports the required libraries 'NLTK' for natural language processing and 'TKINTER' for creating the GUI.
2. It will download the required NLTK resources for tokenization and POS Tagging.
3. To perform the POS tagging a function it will defines as 'tag _text ()' for input text.
4. After it creates the main window('root') for the GUI.
5. Adds a text input field for entering the text to be tagged.
6. To trigger the tagging process it adds a button ('Tag Text').
7. It will add a text area to display the tagged text.
8. For navigating it adds a scrollbar for tagged text.
9. Last it runs the GUI using 'root. Main loop()'.

3. Result

Statistical POS Tagger that we have developed assigns concerned tags to sanskrit text. Firstly plain sanskrit text is taken as input to the statistical model .Then tokenization and feature extraction tasks are completed.The model assigns POS tags with the help of statistical tagger and the ouput is given in the form of tagged sanskrit text.

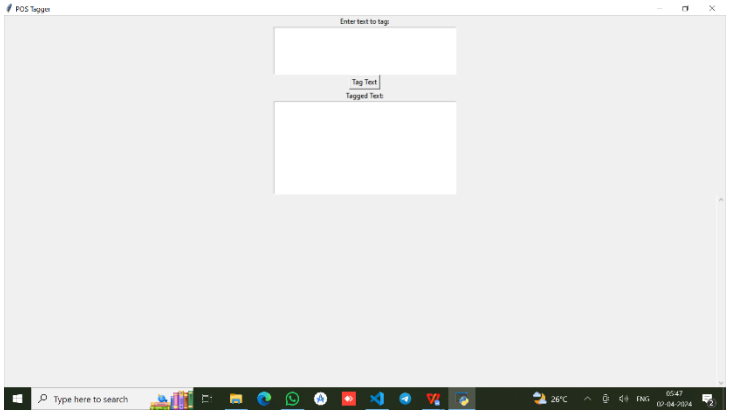

Fig.2 Home Page

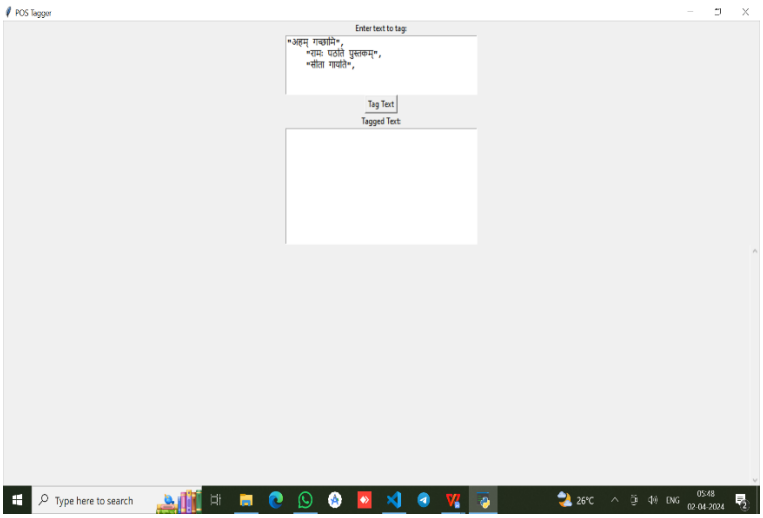

Fig.3 input

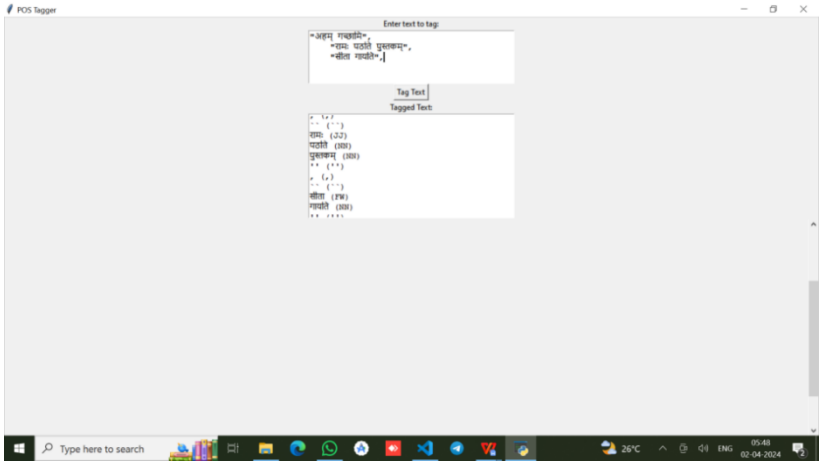

Fig.4 output

4. Conclusion

Statistical part-of-speech (POS) annotation for Sanskrit offers a blend of challenges and prospects. While Sanskrit's distinct attributes, such as its intricate morphology and fluid word arrangement, can present obstacles for precise tagging, statistical models exhibit potential in achieving notable accuracy levels. Researchers have curated specialized datasets, models, and assessment criteria to gauge the efficacy of POS annotation in Sanskrit texts.

On the whole, statistical POS annotation for Sanskrit holds promise in enriching natural language processing Endeavours, including machine translation, information retrieval, and sentiment analysis, by furnishing vital linguistic insights into word classifications and connections. Sustained exploration and advancements in this domain are critical to refining the precision, efficacy, and adaptability of statistical POS annotation models for Sanskrit, paving the way for more advanced and efficient language processing applications in this ancient and intricate language.

REFERENCES:

1. Title: "A Statistical Part-of-Speech Tagger for Sanskrit"Authors: Girish Nath Jha, Dipti Misra Sharma, and Kavi Mahesh Published in: Proceedings of the Workshop on Computational Linguistics for South-Asian Languages, 2009
2. Authors: A. Sharma, A. Mohankumar, et al.Published in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational
3. Authors: Lavanya, K. V. and Ramesh Babu, Ganapathi RajuPublished in: International Journal of Computer Applications (0975 – 8887)
4. Summary: This research paper discusses statistical approaches for POS tagging in Sanskrit, focusing on methodologies and challenges specific to Sanskrit language processing.