INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Assamese Inflectional Rule-Based Stemmer

Mandira Neog[1]*, **Nomi Baruah**[2]

1 Research Scholar, Centre for Computer Science and Application, Dibrugarh University, Assam, India
2 Associate Professor, Department of Computer Science & Engineering, DUIET, Dibrugarh University, Assam, India

OPEN ACCESS

Received: 23-03-2023

Accepted: 26-06-2023

Published: 02-11-2023

Editor: Guest Editors: Dr. Madhuryya Saikia & Dr. Niranjan Bora

Citation: Neog M, Baruah N (2023) Assamese Inflectional Rule-Based Stemmer. Indian Journal of Science and Technology 16(SP2): 38-43. https://doi.org/10.17485/IJST/v16iSP2.6447

* **Corresponding author**.

mandira.neog@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

The research paper presents an Assamese Inflectional Rule-Based Stemmer with supervised rules that tackles the issues brought on by the language's complex morphology. **Objectives:** The goal is to create a stemmer that accurately captures a word's context while taking its inflectional forms into account. **Methods:** This was accomplished using a rule-based approach that had been improved by the incorporation of Parts-Of-Speech tagged (POS-tagged) data with Assamese WordNet. A corpus of 50,000 words from diverse sources was used, together with POS-tagged data. Assamese WordNet was also added to improve the stemmer's performance. **Findings:** An accuracy of 91% achieved utilising the tagged data, surpassing the 86% accuracy attained without the use of tags, the evaluation findings showed a substantial improvement. The suggested stemmer successfully captures word meanings and their inflectional variations by combining POS tagging and utilizing Assamese WordNet. **Novelty:** Applications like sentiment analysis and information retrieval systems benefit from this research's advancement of Assamese language processing. The development of accurate stemming methods successfully closes the prevailing gap in elemental processing. This development enables information retrieval systems to operate more quickly and accurately.

Keywords: Stemming; POStag; NLP; Suffix; Prefix

1 Introduction

For information retrieval, machine translation, and document classification in NLP (Natural Language Processing), stemming [1] is vital. Due to its complex morphology and limited digitally available linguistic resources, Assamese presents particular difficulties for stemming. In order to increase precision and effectiveness in language processing tasks and to boost Assamese language processing accuracy, this study suggests a novel Assamese stemming approach. Assamese stemming techniques that are currently state- of-the-art have explored a number of different strategies, including [2] rule-based, [3] statistical, [4] unsupervised and [5] hybrid methods.

In solving the difficulties of Assamese stemming, each strategy has its own benefits and drawbacks. A common method used in Assamese language processing is rule- based stemming. The earlier research on rule-based stemming systems provided several strategies in various languages. For instance, [6] Gogoi et al. presented a rule-based technique for the Assamese language stemming that focused on noun and verb inflections. This strategy entailed deleting suffixes from words depending on certain principles, allowing the extraction of base forms or stems. [7] Mubasher et al. presented a hybrid stemmer that combines rule-based and LSTM sequence-to-sequence techniques. It did not, however, address contextual word understanding or ambiguous word stemming.

In contrast to the previously reported rule-based stemmers, which were primarily concerned with matching suffixes and prefixes of certain forms, our suggested technique makes use of POS-tagged data. This enables us to address all types of root words based on their POS tags. To reliably strip matching suffixes, we created distinct lists of suffixes and prefixes and wrote specific code for each POS tag. By offering vital contextual information about the grammatical function and syntactic context of words, POS-tagged data improves our stemmer. As a result, our decision-making when recognising and removing suffixes is better informed, resulting in more precise stemming and retention of the words' intended meaning. Our approach incorporates POS tagging to address the challenges of over- and under-stemming associated with ambiguous words, resulting in more accurate stemming outcomes. The major goal of the research is to create an improved Assamese stemmer that, by using effective techniques from Indian languages, maintains the contextual meaning of inflected words in the intricate Assamese language. We want to solve the difficulties brought on by large vocabularies, frequent lexical changes, a lack of language resources, and the need for adaptation. We have developed an enhanced Assamese stemmer using proven Indian language stemming methods to boost language processing skills and enable a variety of applications for Assamese information retrieval and natural language understanding. By integrating contextual analysis and semantic comprehension to handle ambiguity and derive precise root forms, increasing the vocabulary and morphological rules for improved coverage and accuracy, and deftly handling new words and dialectal variances, our technique addresses significant difficulties. The model's performance and linguistic resources are strengthened by utilising POS-tagged data with Assamese WordNet for better information retrieval and natural language interpretation.

2 Methodology

Based on existing stemming techniques, it has been observed that rule-based approaches have shown promising results for Indian languages [2], which predominantly follow the Subject-Object-Verb (SOV) language structure. The application of rule-based techniques in stemming algorithms has demonstrated improved outcomes, specifically tailored to the unique linguistic characteristics of Indian languages. Drawing upon this understanding, this research aims to leverage the advantages of rule-based stemming to develop an effective approach for stemming in Assamese, an Indo- Aryan language spoken in northeastern India.

In the following section, a concise overview of the operational principles of a rule-based stemming algorithm is discussed.

2.1 Rule-Based Approach

To determine the root or stem form of words in a given language, the rule-based approach [6] to stemming employs a methodical procedure. It entails using a predetermined set of linguistic principles to remove affixes from words and produce their base form. The rules were created using the language's morphology, grammar, and orthographic patterns. The following steps are involved in the operation of a rule-based stemming method:

- ***Rule Definition*** **:** Linguistic experts analyze the language's structure and define a set of rules that capture the patterns of affixation and inflection. These rules are typically based on linguistic principles, linguistic resources, and language-specific considerations.
- ***Tokenization*** **:** Tokenization divides the incoming text into discrete words, or tokens. In this step, the text is divided into digestible chunks for later analysis and stemming.
- ***Rule Application*** **:** The given rules are applied progressively to each word or token in order to establish the stem form. The rules recognise affixes, like prefixes and suffixes, and eliminate them in accordance with particular circumstances and grammatical patterns.
- ***Stem Extraction*** **:** The final form is regarded as the word's stem or root form once the rules have been followed. The stem could go through additional normalisation steps, including being made lowercase or getting linguistic adjustments made.
- ***Stemming Output*** **:** The final product of the rule-based stemming procedure is a group of stemmed words that represent the original words' root forms. These stems can be applied to a variety of NLP tasks, including text classification, language

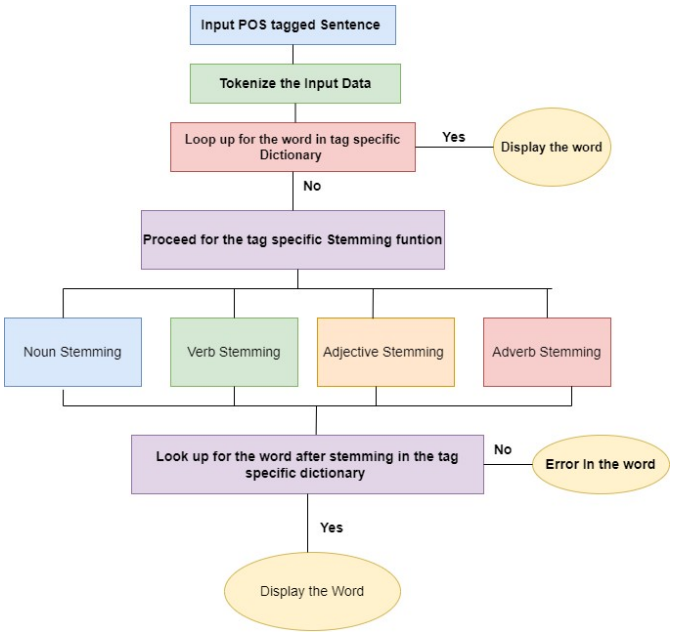

Fig 1. Pictorial diagram of the proposed Assamese Stemmer

analysis, and information retrieval.

There are various benefits to the rule-based stemming method used in SOV-structured languages, including many Indian languages. First of all, it enables the creation of rules that are unique to the language and fit with its morphology, grammar, and orthographic patterns. Better stem extraction accuracy is guaranteed by this customised strategy. Second, the adaptability and control offered by the rule-based method allow the stemming procedure to be adjusted to handle different inflectional forms, derivational variations, and language-specific exceptions.

The interpretability of rule-based stemming algorithms additionally enables simple rule modification and adaptation to particular linguistic variants or dialectal forms. By capturing the morphological patterns and utilising language-specific rules, the rule-based technique is computationally effective, needs little in the way of computational resources, and can produce accurate stem extraction findings. In light of these benefits, the rule-based approach is a useful method for efficiently processing and analysing text in SOV-structured languages.

2.2 Dataset Presentation

With a rich linguistic history and a free word order, Assamese allows for every possible arrangement of the subject, object, and verb inside a phrase, including SOV, SVO, VSO, OVS, and OSV. The parts of speech in a language must therefore be thoroughly understood in order to understand language structure. Adjectives, verbs, nouns, and adverbs all fall under the open class and have a variety of inflectional forms. Our objective is to locate the word's root. In the worst cases, a noun can have more than 25,000 inflected forms, making nouns and verbs the two main problems. We are using a new Assamese dataset with as many Assamese-inflected terms as we're able to find. The experiment was run at the sentence level with appropriately labeled data, preserving the maximum amount of contextual Assamese language meaning. The data was manually gathered from various Assamese articles and Newspapers etc. The gathered data is then transformed into a CSV file using the encoding type UTF-16 LE, which allows Assamese script when converting to CSV format. Following are the few samples of our dataset where we move suffixes to get the base words we need. Due to the fact that just suffixes are searched for, the process is quick. We've provided a list of various location- specific prefixes, including বাৰী, নগৰ ,পাৰা. Additionally, as can be seen, some suffixes are applicable to both place names and human names: ৰ,ক,ৈল.

Table 1. Example of Assamese Suffix

Root	Surfaceform	Suffix
অসম	অসমীয়া	◌◌ীয়া
◌েতজপুৰ	◌েতজপুৰীয়া	◌◌ীয়া
ভাৰত	ভাৰতীয়	◌◌ীয়
অসম	অসমবাসী◌েয়	বাসী◌েয়
ভাৰত	ভাৰতৰত্ব	ত্ব

2.3 Proposed Approach

Although they use the same technology, they include various rule-based methods into the algorithm that best suit their word structures and dialects. For instance,[8] Kariyawasam et al. created a rule-based stemming method for the Sinhalese language in a prior work that relied on suffix and prefix matching. This approach, however, was restricted to the use of conventional prefix and suffix stripping methods.[9] Shakib et al. focused on verb and number inflections in stemming Bangla words in another study, but they did not completely utilise Parts of Speech (POS) labelling, which resulted in the loss of ambiguous word forms.[6] Gogoi et al. suggested a rule-based suffix stripping method for Assamese that focused especially on noun and verb inflections. Another method by Sarmahet al.[10] made use of WordNet to power a dictionary lookup and rule-based stemmer. Assamese language stemming algorithms are discussed in[6] and[10].

Rule-based suffix stripping algorithms are common approach used in these methods. This eliminates the need for a look up table that contains various inflected forms of word and root form relations in order to match words. In this instance, a particular rule is specified to provide the best justification for the Assamese language and word structure. There are many words with different inflections in Assamese because it is one of the languages in India with the maximum morphological diversity. It is quite challenging to understand the context-specific meaning of Assamese due to the variability in spelling of the stem words, which is the main linguistic problem with Assamese stemming. In an effort to enhance the outcomes, we are using a new Assamese dataset with as many Assamese-inflected terms as we're able to find. The experiment was run at the sentence level with appropriately labeled data, preserving the maximum amount of contextual Assamese language meaning.

Following are the different stages involved in the development of our Inflectional Rule-Based Stemmer:

- Input Assamese text data with POS tags.
- Tokenize the input data.
- Look up the given word in the tag-specific Dictionary using its specified tag.
- If the word is not found in the tag-specific dictionary, the next step is to match the suffix of the word with the specified suffix list. If a match is found, Rule-based Suffix Stripping is carried out.
- Loop-up for the new word obtained after rule-based suffix stripping in the tag-specific Dictionary. If a matching word is found, it will be displayed.
- If the word cannot be found, the previously specified word (word before stripping) is obtained. Then look up for the word in the inflectional word list to see if it has any word variations.
- If the word is found, it is replaced with its original word using Assamese WordNet.

3 Results and Discussion

The database for this experiment was manually built, and all of the data were tagged using an Assamese tagger before being fed into the model. In order to evaluate the effectiveness of the suggested model and the effectiveness of the suggested technique, we have gathered as many terms as we could with ambiguous meanings. 25% of the data are utilized to test the model, while the remaining 75% are used for training. Using Python, we created a system for Assamese Stemmer based on the afore mentioned algorithm. In order to assess the system's performance, we have calculated accuracy. For the Information Retrieval System to function at its best in Assamese, it is important to comprehend the word's context. Assamese language contains a large number of inflection variation terms. As a result, knowledge of the language and its word structure is crucial for understanding a word's precise meaning when it is used in a complex context peculiar to that language. Through the use of tagged data, the model's complexity has been reduced, and system performance has increased.

Our suggested system works by removing suffixes depending on a word's unique tag. To extract the root word for each tag, such as a noun, verb, adverb, adjective, etc., a different piece of code is written for each one.

We have used the following equation to evaluate the accuracy of our system:
Accuracy = $\frac{TN+TP}{TN+FP+TP+FN}$
Where,
TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative

According to our results analysis, using tagged data improves the system's performance. Our main goal was to eliminate the language complexity challenges faced by other available Assamese Stemmers, as described above. As a result, we designed our algorithm to account for all of the limitations and used Assamese tagged data to achieve a better result. Table 2 compares the outcomes produced by our suggested stemming model with the available Assamese Stemmer. In reference[10], a look-up and rule-based suffix stripping approach based on WordNet obtained 85% accuracy for Assamese stemming. The stemmer's validation required comparing hamming distances with morphological root words, although it was confined to Assamese WordNet and Named Entities dictionary entries. Reference[6], on the other hand, proposed a rule-based stemming approach that was limited to noun and verb word inflections. This work proposes a novel inflectional rule-based Assamese stemmer that uses Assamese WordNet and performs suffix stripping depending on the POS tag of each word. This method offers an advantage and is an improvement over existing methods.

Table 2. Comparison of stemming results with other Assamese Stemmer

Language	References	Data	Accuracy
Assamese	[6]	20,000 words	86.16%
Assamese	[10]	2 Lakh	85%
Assamese	[proposed model]	50,000 words	91%

The study of the results reveals that the range of other Assamese stemmers was restricted to noun and verb word inflections. The dataset's size was also restricted in order to produce more encouraging results. Both of these issues have been solved in our suggested stemmer. We have a larger dataset with a greater proportion of Assamese inflectional vocabulary than the other Assamese stemmers. The calculated results of the suggested system are assessed by taking into account 50,000 words. Stemming of the Assamese words are performed based on four parts-of-speech (POS) tags, such as noun, verb, adverb, and adjective. Since POS tags provide a more thorough, intricate contextual understanding of the language, this help us achieve a promising accuracy of 91%, which is quite better than the other available Assamese stemmers.

Table 3. Comparison of stemming results with tagged data and without tagged data

	Accuracy	
Language	Tagged Data	Untagged Data
Assamese	91%	86%

The Table 3 compares the results obtained by our proposed Assamese Inflectional Rule-Based Stemmer when using tagged data and without using tagged data. The accuracy of 91% is achieved when the stemmer is used with tagged data. This is 5% better than the results obtained without using tagged data, which had an accuracy of 86%.

4 Conclusion

The study presents an improved Assamese stemmer that preserves the contextual meaning of inflected words in the complex Assamese language. The stemmer achieves exceptional accuracy by using a rule-based technique with POS tag data and Assamese WordNet. On a dataset of 50,000 words, the evaluation shows a remarkable 91% accuracy with tagged data, outperforming the 86% accuracy attained without tags. The suggested stemmer outperforms prior techniques, attaining 85% accuracy using a WordNet-based approach confined to Assamese WordNet and Named Entities and concentrating primarily on noun and verb inflections.

The utilisation of POS tagged data, which permits the retention of contextual meaning and efficient handling of ambiguous words, is the core strength of our research. The small size of the dataset used, however, raises the possibility of a weakness and points to the requirement for dataset extension. Furthermore, it's yet unclear how difficult it will be to get precise POS-tagged data from user-generated content. The study's significance lies in providing a more accurate stemmer for handling inflected

Assamese terminology, with potential for further comparisons and improvements. Future research can explore expanding the data size and developing an Assamese lemmatizer to enhance the stemmer's capabilities.

5 Declaration

Presented in Fourth Industrial Revolution and Higher Education (FIRHE 2023) during 23rd-25th Feb 2023, organized by DUIET, Dibrugarh University, India. The Organizers claim the peer review responsibility.

References

1) Jabbar A, Iqbal S, Tamimy MI, Hussain S, Akhunzada A. Empirical evaluation and study of text stemming algorithms. *Artificial Intelligence Review volume*. 2020;53:5559–5588. Available from: https://doi.org/10.1007/s10462-020-09828-3.
2) Sharipov M, Yuldashov O. UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. In: The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP), Koper, Slovenia. 2022. Available from: https://doi.org/10.48550/arXiv.2210.16011.
3) Singh P, Bhowmick PK. Neural Network Guided Fast and Efficient Query-Based Stemming by Predicting Term Co-occurrence Statistics. *SN Computer Science*. 2022;3(198). Available from: https://doi.org/10.1007/s42979-022-01081-5.
4) Nathani B, Joshi N, Purohit GN. Design and Development of Unsupervised Stemmer for Sindhi Language. *Procedia Computer Science*. 2020;167:1920–1927. Available from: https://doi.org/10.1016/j.procs.2020.03.212.
5) Alobed M, Altrad AMM, Bakar ZBA, Zamin N. Automated Arabic Essay Scoring Based on Hybrid Stemming with Wordnet. *Malaysian Journal of Computer Science*. 2021;(Special Issue 2):55–67. Available from: https://doi.org/10.22452/mjcs.sp2021no2.4.
6) Gogoi A, Baruah N, Sarma SK, Phukan RD. Improving stemming for Assamese information retrieval. *International Journal of Information Technology*. 2021;13:1763–1768. Available from: https://doi.org/10.1007/s41870-021-00718-7.
7) Malik MH, Ghous H, Ahsan I, Ismail M. Saraiki Language Hybrid Stemmer Using Rule-Based and LSTM-Based Sequence-To-Sequence Model Approach. *Innovative Computing Review*. 2022;2(2):18–40. Available from: https://doi.org/10.32350/icr.0202.02.
8) Kariyawasam KTPM, Senanayake SY, Haddela PS. A Rule Based Stemmer for Sinhala Language. In: 2019 14th Conference on Industrial and Information Systems (ICIIS), 18-20 December 2019, Kandy, Sri Lanka. IEEE. 2019. Available from: https://doi.org/10.1109/ICIIS47346.2019.9063286.
9) Shakib MSS, Ahmed T, Hasan KMA. Designing a Bangla Stemmer using rule based approach. In: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), 27-28 September 2019, Sylhet, Bangladesh. IEEE. 2019. Available from: https://doi.org/10.1109/ICBSLP47725.2019.201533.
10) Sarmah J, Sarma SK, Barman AK. Development of Assamese Rule based Stemmer Using WordNet. In: Proceedings of the 10th Global Wordnet Conference. Global Wordnet Association. 2019;p. 135–139. Available from: https://aclanthology.org/2019.gwc-1.17.