# Development of Assamese Rule based Stemmer using WordNet

**Anup Kumar Barman**
Dept. of IT
Central Institute of Technology
Kokrajhar, India
ak.barman@cit.ac.in

**Jumi Sarmah**
Dept. of IT
Gauhati University
Guwahati, India
jumis884@gmail.com

**Shikhar Kr. Sarma**
Dept. of IT
Gauhati University
Guwahati, India
sks001@gmail.com

## Abstract

Stemming is a technique that reduces any inflected word to its root form. Assamese is a morphologically rich, scheduled Indian language. There are various forms of suffixes applied to a word in various contexts. Such inflected words if normalized will help improve the performance of various Natural Language Processing applications. This paper basically tries to develop a Look-up and rule-based suffix stripping approach for the Assamese language using WordNet. The authors prepare the dictionary with the root words extracted from Assamese WordNet and Named Entities. Appropriate stemming rules for the inflected nouns, verbs have been set to the rule engine and later tested the stemmed output with the morphological root words of Assamese WordNet and Named Entities by computing hamming distance. This developed stemmer for the Assamese language achieves accuracy of 85%. Also, the authors reported the IR system's performance on applying the Assamese stemmer and proved its efficiency by retrieving sense oriented results based on the fired query. Thus, Morphological Analyzer will embark the research wing for developing various Assamese NLP applications.

## 1 Introduction

Computationally, stemming is the process to automatically extract the base form of a given inflected word. The stemmed word is not required to be identical with the morphological root of the word. Most Indian languages are highly inflectional and many words in a document appear in many morphological forms. Indexing is the important subtask of an IR system. Indexing all words in a document appearing in various morphological forms is highly tedious and time-consuming. Thus, it is necessary to stem the words to reduce them to their original base form. Reducing to their original base form will help the indexer in IR to detect the important terms in a document, detect Named entities, multi-word expression and extract stopwords. Looking deeply into the matter, we found that two parts-of-speech Nouns and Verbs have a wide list of inflections for the Assamese language. The main objective of this paper is to perform stemming task on a group of inflected words to retrieve root words with an acceptable accuracy.

Many approaches to stemming have been identified. They are classified into three categories- Rule-based, Statistical and Hybrid approaches.

*Rule-based approach*- Such approaches apply a set of morphotactic rules of a language to an inflected word. Such rules may derive the base form by emitting the suffix or the prefix.

*Statistical approach*- One of the drawbacks of rule- based approach is that it is language dependent and it is dependent on the database. Statistical approach overcomes both the problems by calculating probabilistic distributions of the terms.

*Hybrid approach*- Combination of both rule-based and statistical approaches.

In this paper, the authors have researched and implemented a rule-based stemmer for Assamese language embedding the Look-up based approach. The quick Look-up approach is made on the dictionary prepared from Assamese WordNet and Named Entities. Assamese WordNet is a large lexical knowledge database developed by the team (Sarma et al., 2010). It contains four major components-

- ID: an unique identification number
- CAT: the Parts-Of-Speech category
- Synsets: the main building block of WordNet. A number of 30K synsets are present in

Assamese WordNet

- Gloss: The concept or meaning of the given synset

Named entities are a collection of terms that has a unique concept. They are mainly the names of people, organization, places, festivals etc.

Assamese is the official language of the North-eastern state- Assam of India. It is spoken by nearly 15 million people. Assam shares an international border with Bhutan and Bangladesh. It is a computationally less aware language which belongs to the Indo-Aryan language family. But, recently some development is done for this language from Natural language processing perspective. Development of Assamese WordNet, Corpus, IR system is some of them.

This research paper aims to implement a rule-based Morphological Analyser for the Assamese language to be embedded as a plug-in to Assamese IR system. No such work implementing 22 morphotactic rules for Assamese language is defined before in previous works. We believe this would mark a great contribution to Assamese NLP area.

The road-map of the paper is as follows- Section 2 discusses some related work to stemming implemented in Indian languages, Section 3 describes the rule based stemmer for the Assamese language with the system architecture. Section 4 discusses the performance of the stemmer computing the hamming distance. The IR system performance is evaluated on performing stemming to the inflected terms and the results are reported in section 5 of this paper. The paper is summarized in Section 6.

## 2 Background work

This section gives us an overview of stemmers developed in Indian Languages. For the English language, the most commonly used stemming algorithm is the Porter stemming algorithm (Willett, 2006) which followed a rule-based approach. The Indian language (Ramanathan and Rao, 2003; Aswani and Gaizauskas, 2010; Mahmud et al., 2014; Kumar and Rana, 2010; Majgaonker and Siddiqui, 2010; Prajitha et al., 2013; Thangarasu and Manavalan, 2013; Kumar et al., 2011) in which stemmer is developed along with the approaches used and accuracies derived is mentioned in Table 1

Table 1: Indian language stemmer

| Language | Approache | Correctness |
|---|---|---|
| Hindi | Rule-based | Accuracy 88% |
| Gujarati | Dictionary and Rule-based | Precision 83% |
| Bengali | Rule-based | Accuracy 88% |
| Punjabi | Brute-force | Accuracy 81.27% |
| Marathi | Hybrid (Rule-based + suffix stripping +statistical) | Precision 82.50% |
| Malayalam | Finite state machines | Accuracy 94.76% |
| Tamil | Light Stemmer (preserves word meaning) | Accuracy 83.28% |
| Telugu | Unsupervised approach | Accuracy 85.40% |

## 3 Development of Assamese stemmer

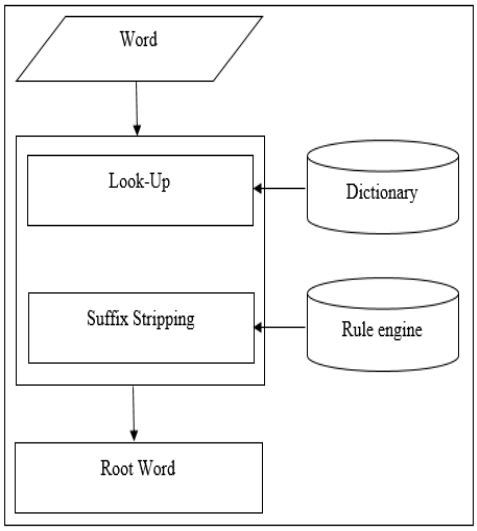


Figure 1: Assamese stemmer system diagram

Assamese words in a text take a series of suffixes in a sequential manner. For developing a rule-based stemmer, our first intention is to determine the sequence of various suffixes a word can occur in a text. Some of them were collected by consulting with the Linguistic scholars of GU NLP team. They may be divided into eight possible suffix categories such as:

- Plural- "সকল", "মখা", "বিলাক", "সোপা"
- Case markers- "ক", "ত", "লৈ", "এ", "ৰ"
- Pleonastic suffix- "হে", "চোন", "নে", “গৈ”
- Definitive- "জন", "জনী", "জনা", "খন", "টো"
- In-definitive- "কেইজন", "কেইজনী”
- Verbal- "াই", "াইছ", "াইছা", "াইছিল"
- Kinship noun- “য়েক”
- Extra- “দৰে”,"য়ে", "নো", "কৈ"

**Step1: Dictionary Lookup**

Assamese dictionary of size about 2 lakh root words is prepared by our Linguists from Assamese WordNet and Named Entities. Our module first looks at the dictionary table to determine if the words are already in the root form. If true then, they proceed to step 3 else step 2. This approach eliminates the type of error like word saybাহিৰ (out), which is a root word even though case marker suffix is present. If the dictionary is not reviewed in the beginning, than stemmer would remove the suffix of the word which would lead to overstemming. Moreover, the same would be the case for Named Entities like place name: তেজপুৰ (place name). Also, in some cases the term may have been derived from the antonym of the root word. Here, we consider the antonyms as the root word to retrieve sense oriented searched results from an IR. As for example the word in Assamese language- অশুভ (not pleasant) indicates different sense compared to the root form শুভ (pleasant). On knowing the root words at beginning will avoid understemming and overstemming roles of the stemmer and can retrieve sense oriented or meaningful results from the Information retrieval system on firing the query as required by the user.

**Step2: Suffix pruning**

If the first step fails than step 2 is executed. In this phase, the rule engine generates a list of suffixes in a proper manner that may be attached to the root based on the stemming rules already incorporated in the engine. The generated suffix list must abide by the morphotactic rules for Assamese. A Java program was developed to run this step.

Some rules for stemming are mentioned below in a tabular form: Here, authors have defined 22 rules for stemming Assamese words. Some of the rules are followed by the Assamese grammar book Assamiya Vyakaran by Hem Chandra Baruwa, 2003. As for example, the inflected word is মানুহকেইজনমান. The generated suffix list for the word is মান, কেইজনমান. The list is now transformed to non-increasing order and at first the top one (here কেইজনমান) is being tried to be matched with the already incorporated rules in the engine. Here, the rule *root+ indefinitive + plural* is mapped and the word is stemmed. Here, at the first phase of developing the stemmer, only nouns and verbs are taken into consideration.

Table 2: Morphotactic Rules of Assamese Stemmer

| **Suffix Type** | **Assamese Notation** |
|---|---|
| Root+casemarker | মানুহ+ৰ |
| Root+definitive | মানুহ+জন |
| Root+pleonastic | কৰ+চোন |
| Root+indefinitive | মানুহ+কেইজন |
| Root+plural | মানুহ+বোৰ |
| Root+verb | কৰ+ িছিল |
| Root+extra | খৰ+কৈ |
| Root+kinshipnoun | ককা+য়েক |
| Root+case+extra | মানুহ+ৰ+দৰে |
| Root+plural+ case+pleo | মানুহ+বোৰ+ ক+হে |
| Root+Plural+ Case marker | মানুহ+বোৰ+ৰ |
| Root+Plural+ pleonastic | মানুহ+বোৰ+হে |
| Root+Definitive+ case | মানুহ+জন+ৰ |
| Root+Definitive+ pleonastic | মানুহ+জন+হে |
| Root+Indefinitive+ Plural | মানুহ+কেইজন+ মান |
| Root+Verb+ pleonastic | পঢ়+ ইলে +গৈ |
| Root+Casemarker + pleonastic | কৰ+ক+চোন |
| Root+kinshipnoun+ indefinitive+plural+pleo | নাতিনী+য়েক+ কৈইজনী+মান+হে |
| Root+pleonastic+ pleonastic | কৰ+গৈ+চোন |
| Root+plural+ definitive | গৰু+জাক+টো |
| Root+verb+extra | কৰ+ি+য়ে |
| Root+case+ plural+definitive | গৰু+ৰ+জাক+টো |

**Step3: Exit**

## 4 Performance Analysis

We have implemented both look-up based and rule-based approaches for Assamese stemmer. We evaluated the stemmed output with the morphological root words of Assamese WordNet and Named Entities by computing Hamming distance. It is the number of different position of the bits between two equal length strings. A hamming distance of 0 means the two strings are equal in both position of the character bits and weight. As for example one of the correctly stemmed output is:

Inflected term: মানুহজন

Assamese Stemmer output: w1= মানুহ

Assamese WordNet (ID: 196) w2= মানুহ

Hamming distance=d(w1,w2) =0

Some of the result statistics found while analyzing the performance of the stemmer is shown in a tabular form below:

Table 3: Statistics of stemmer performance

| | |
|---|---|
| Correctly stemmed | 85% |
| Incorrectly stemmed | 15% |

## 5 Stemmer in Assamese IR

Information Retrieval system retrieves relevant and sense oriented information to a user based on the query. Assamese NLP aims to develop a monolingual search engine which will help the web users to retrieve information in ones own native language say Assamese. Only a few (2-3) percent of people of Assamese community knows to speak, read or write English, so retrieving information in own language will be much benefited.

Assamese IR system is technically composed of two parts- Apache Solr & Nutch. Apache Solr is an open source search platform written in JAVA from Apache Lucene project. Some of the major features of Solr are- full text search, real time indexing, dynamic clustering etc. Apache Nutch is also a JAVA coded tool with the crawler feature. The crawler can be biased to fetch important relevant pages at first. We developed Assamese monolingual system considering Solr3.4 and Nutch1.4 as indexer and crawler respectively.

Stemming is an important plug-in of IR. Stemming is performed to an inflected word to avoid mismatches between words that share the same root word. Let us consider a simple example- if we are searching for a document entitled Ways to write a book and the user issues a query writing, than there will be no match with the title. But, if the query is stemmed before than the search system will stem the word writing to write and the retrieval will become easier and successful. Stemming is applied to both Query processing module and IR system module. Both at the indexing time and during processing of the query the stemmer module is added as plug-in to Assamese IR system. Here, we have analyzed the performance of IR system based on two categories-

- IR performance without stemming
- IR performance with stemming

The above two techniques is evaluated with p@k (Precision at k) metric. For modern IR system, recall is meaningless as many numbers of queries retrieves many relevant documents (as of now web-scale) and no user will go through all of them. Here, k=10 and p@10 indicates the number of relevant result of search result page which includes top-ten results of a query. To evaluate our system we tokenized some of the words from Assamese Corpus (size=1.5 million words) developed by (Sarma et al., 2012). The figure 2 indicates higher AP (Average Precision) values of the IR system when performed stemming than without stemming. To evaluate the system we consider 10 Assamese queries Q1 to Q10 those are অমৃতসৰৰ স্বৰ্ণ মন্দিৰ, তিৰুপতি, নালন্দা বিশ্ববিদ্যালয়, কাজিৰঙা ৰাষ্ট্রীয় উদ্যানলৈ, বিহুত, অসমৰ, মাজুলী, তাজমহলত, গড়ঁবোৰ ,

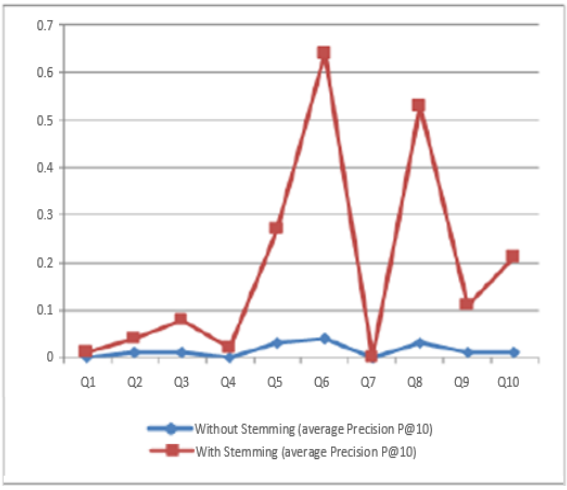


Figure 2: Assamese IR performance: with and without stemming

ব্রহ্মপুত্র নদীত. As the stemmed term indicates larger concept than the original term appears in the document, the stemming increases the number of retrieved relevant documents.

## 6 Conclusions

The performance of the Assamese stemmer mentioned in this paper shows that it attains a state of art accuracy as a stand along system as well as a component of Information Retrieval system. The proposed technique is Dictionary Look-up and Rule-based approach for this Indo-Aryan language with an acceptable accuracy of 85% and 22 defined morphotactic rules.Increasing the dictionary size will result in more increasing accuracy.

Assamese stemmer is the basic language resource and is used in many applications in the field of Text mining and NLP like IR, MT, Document Classification etc. The accuracy of the stemmer can be improved by defining more stemming rules and increasing the dictionary size with more root words. Moreover, as the IR performance on performing stemming to the inflected terms indicates an overwhelming result, thus stemmer is an important resource for Assamese NLP.

## References


Shikhar Kr. Sarma, Moromi Gogoi, Utpal Saikia and Rakesh Medhi 2010. *Foundation and structure of Developing Assamese WordNet*. In Proceedings of 5th International Conference of the Global WordNet Association.

P. Willett 2006. *The Porter stemming algorithm: then and now*. Program: electronic library and information systems, 40 (3).

A. Ramanathan and D. D. Rao 2003. *A Lightweight Stemmer for Hindi*. Workshop on Computational Linguistics for South-Asian Languages, EACL.

N Aswani, R Gaizauskas 2010. *Developing Morphological Analysers for South Asian languages. Experimenting with the Hindi and Gujarati languages*. In Proceedings of the seventh conference on International Language resources and evaluation, Malta.

Md. Redowan Mahmud, Mahbuba Afrin, Md. Abdur Razzaque, Ellis Miller and Joel Iwashige 2014. *A rule based Bengali stemmer*. In Proceedings of the ICACCI.

D Kumar and P Rana 2010. *Design and development of a stemmer for punjabi*. International Journal of Computer Application 11(12) (December 2010).

M. M. Majgaonker and T. J. Siddiqui 2010. *Discovering suffixes- A case study for Marathi language* . International Journal on Computer science and engineering.

U. Prajitha, C. Sreejith and P.C Reghu Raj 2013. *LALITHA: A light Weight Malayalam Stemmer using Suffix Stripping method*. In Proceedings of the ICCC.

M Thangarasu and Dr R Manavalan 2013. *Stemmers for Tamil Language: Performance Analysis*. International Journal Of Computer Science & Engineering Technology, Vol 4.

A . P. Siva Kumar, P. Premchand and A Govardhan 2013. *TelStem: An Unsupervised Telugu Stemmer with Heuristic Improvements and Normalized Signatures*. Journal of Computational Linguistics Research Vol 2 number 1.

Shikhar Kr. Sarma, Himadri Bharali, Ambeswar Gogoi, Ratul Ch. Deka, Anup Kr. Barman 2012. *A Structured Approach for Building Assamese Corpus: Insights Applications and Challenges*. In Proceedings of the 10th Workshop on Asian Language Resources, pp. 21-28, December.