Design of an Inflectional Rule-Based Assamese Stemmer

Swagata Seal, Nisheeth Joshi

***Abstract*: Assamese is a very morphologically rich language. A little work has been done on Assamese Language Processing. . As Assamese is one of the most resource poor languages in the field of computational studies thus, we intend to present an inflectional rule-based stemmer for Assamese language. Stemming is the simplest and prior step for natural language processing (NLP), it is a procedure which removes the suffixes from the root word. This performs very little morphological analysis. After stemming the resultant word is known as 'Stem' or root word. The proposed system is language dependent and domain independent. A suffix stripping algorithm is used to design the system. The system is evaluated with 20,000 words.**

Index Terms: Assamese, resource poor language, Stemming, Suffix stripping.

I. INTRODUCTION

Morphological analysis is a major level of linguistic analysis. The actual meaning of Morphology is study of the internal architecture of words. Morph means "architecture" and ology means "the study of". Fig 1, shows the example of morphology of a word.

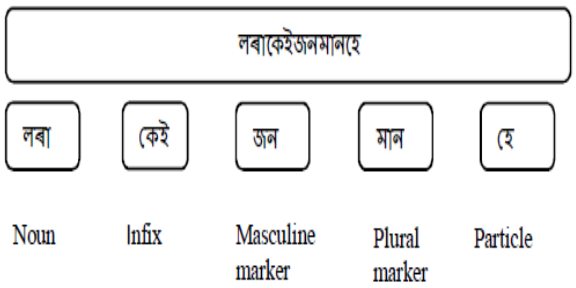

Fig1: Morphology of a word

Two morphological systems are - Inflection and derivation. Inflectional morphological system is making word forms of word and derivation morphological system is construction of new words. Morphemes are the minimal unit of language which bears acceptable meaning; it is the ingredient of Morphology. Morphological analysis means excerpting root word. In the age of NLP we discover such circumstance where more than one word have same root. For example-অসম, অসমীয়া, অসমবাসী | so it is very vital to link these words to its root. To extract root word from these kind of words mainly two techniques Lemmatizing and Stemming are applied. The only difference between Lemmatizing and stemming is lemmatization turns the word into a meaningful form whereas stemming is done by stripping off the affixes from the word.

Stemming is the necessary aspect of search engine, information retrieval system, questionnaire, domain analysis, natural language processing, machine translation and many more. It is a procedure where words are reduced to 'Stem' after applying set of rules or algorithm. It is not compulsory that the stem is an existing word in dictionary, but all its variations must be connected to the stem. As Assamese is one of the most morphologically rich languages thus it is very challenging to determine the stem word, another main problem is Assamese has single letter suffixes which creates ambiguity. For e.g. , root word of the verb খোৱা (eat) has different forms like খাইছিল (ate), খাওক (eat),খাব (eat), খালো (eaten), খালি (eat), খাবি (eat), খাইছো (eaten) etc. Similarly, noun root word বিদ্যালয় (school) has different forms like বিদ্যালয়ত (in school), বিদ্যালয়ৰ (school's), বিদ্যালয়লৈ (to school).

Assamese is spoken by around 20million people; its origin is Sanskrit and that is why it is morphologically rich language and it belongs to the family of Indo-Aryan language. Assamese is widely spoken in Assam, certain sector of Arunachal Pradesh and some other northeastern states also. A very little computational work has been done on Assamese. As it is a resource poor language thus required for development.

II. LITERATURE SURVEY

A number of stemming algorithms are developed, the headmost work was done by Lovins [1] in 1968.She described two major methodology for structuring stemming algorithm i.e., iteration and longest-match in five steps. Thenceforth Martin Porter [2] in 1980 build up an algorithm for removing suffixes, which is a rule-based stemmer followed by five steps. Porter stemmer is considered as one of the best stemmers till date. This algorithm was firstly developed for English language but later on it was developed for some other European languages. Paice [11] proposed another algorithm for stemming named as "Another Stemmer" which performs 4 steps to give final result.

Towards Indian languages Larky [6] worked on Hindi by removing 27 suffixes. Raman than and Rao [3] also worked on the same technique but they used 65 inflection suffixes to remove. The result was achieved by simply taking off the longest possible suffix provided in the list. Majgaonker and Siddiqui [4] in 2010 worked on discovering suffixes for Marathi language. They used two approaches i.e., rule

Revised Manuscript Received on April 19, 2019.

Swagata Seal, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India.
Nisheeth Joshi, Department of Computer Science, Banasthali Vidyapith, Rajasthan, India.

based and unsupervised stemmer. Comparatively unsupervised stemmer gave better result. Shammari and Lin [5] in 2008 presents an error free Arabic. On using Educated Text Stemmer Arabic stemming algorithm 96% correct stem was generated. Ameta et al. [7] proposed a lightweight Guajarati stemmer where by using algorithm longest possible suffix from the list had been removed. Gupta et al [8] designed a stemmer for Urdu language which was rule based, 119 rules were created among which 107 were for postfix and 12 for prefix. Their system was based on stripping the rules. Mahmud et al. [9] created a rule based Bengali stemmer which used hierarchical approach for excerpting stem from all probable verb and noun inflections from a provided word list. Patel and Patel [10] in 2017 introduce a rule-based stemmer using dictionary approach named as "GUJSTER". Accuracy achieved was 97.09%. Jenkins and Smith [12] presented a conservative stemming to stem the correctly spelled words. Its objective was to help in searching and indexing. 85% 0f accuracy was achieved by the system.

In context of Assamese language Saharia el al. [13] worked on stemming algorithm which uses suffix stripping algorithm with rule engine. Accuracy gained was 82%. They also worked on resource poor languages like Assamese, Bengali, Bishnupriya Manipuri and Bodo language [16]. The used HMM based hybrid approach. For Assamese and Bengali accuracy gained was 94%, 87% and 82% for Bishnupriya Manipuri and Bodo. In suffix-based noun and verb classifier [14] noun inflection and verb inflection were used. Sharma et al. [17] proposed a NER (Location Name) in Assamese based on Suffix removing followed by five steps. F-measure of nearly 90% was gained by the proposed system.

III. Structure of Assamese Language

Assamese is a linguistically abundant in nature and is free word order i.e. subject, object and verb can be placed anywhere in a sentence; it can be in the form of SOV, SVO, VSO, OVS, OSV. For e.g. মই (S) স্কুললৈ (O) যাম (V) and স্কুললৈ (o) যাম (v) মই (s) both the sentences have same meaning but can be written differently. Assamese has two genders- পুংলিঙ্গ (Masculine), স্ত্রীলিঙ্গ (Feminine). Open class i.e. noun, verb, adjective and adverb can have multiple inflectional forms. Our aim is to extract the root word. Noun and verb are the major challenge; a noun can have more than 25,000 inflected forms in worst cases.

A. Noun

A noun is simply name of anything i.e. name of person, place, animal and thing. Noun can have Case Marker suffix, Plural Suffix, Gender suffix, Classifier etc. Nouns can be derived from noun stems too, and all suffix rules applied to noun can be applied to pronouns too. Table I though V shows examples noun inflections.

Table I. Noun with case markers

Case marker	Stem	Word
-□	□□□	□□□□
-□□	□□□	□□□□□
-□□□□	□□□	□□□□□□□
-□□□□□□	□□□	□□□□□□□□□

Table II. Plural suffix with noun

Plural suffix	Stem	Word
-বোৰ	মানুহ	মানুহবোৰ
-বিলাক	গৰু	গৰুবিলাক
-হঁত	মা	মাহঁত
-মালা	চিত্র	চিত্রমালা
-জাক	চৰাই	চৰাইজাক
-খোপা	ফুল	ফুলখোপা

Table III. Gender Suffix with Noun

Gender	Stem	Word
-জন (masculine)	□□□□□	□□□□□□□
-জনী (feminine)	□□□□□	□□□□□□□□

Table IV. Classifier Suffixes with noun

Classifier	Stem	Word
-□□	□□□□□	□□□□□□□
-□□	□□□□□	□□□□□□□
-□□□	□□	□□□□□

Table V. Nouns derived from noun stem

Suffix	Noun	Derived Noun
-□□	□□	□□□□
-□□□	□□□	□□□□□□
-□	□□□□□□	□□□□□□□

B. Verb

Suffixes of verbal roots are more complicated than others. Verbs are words which illustrate the activity of a noun. Verb roots are inflected with persons and tense. These two are added to verb root which makes it infinite form. Assamese verb root বহ is reported with 520 inflectional forms [14]. Table VI shows examples of some verb inflections.

Table VI. Inflectional form of verb with respect to person and tense [14]

Inflection with respect	1st person	2nd person	2nd person (respect)	3rd person
Present	খাওঁ	খা	খোৱা	খোৱায়
Past	খালোঁ	খালি	খালে	খালে
Future	খাম	খাবি	খাবা	খাব
Present perfect	খাইছোঁ	খাইছ	খাইছা	খাইছ
Past perfect	খাইছিলোঁ	খাইছিলি	খাইছিলা	খাইছিল
Future Conditional	-	খবিচোন	খাৱাচোন	খাবচোন

C. Adverb

Adverb is the word that restricts the form of a verb. Some questions like how long? where? how much? when? how? are answered by an Adverb. Adverbs can be classified into simple phrasal adverb and complex adverb. At this time our matter of discussion is complex adverb. Table VII shows examples of some adverb inflections.

Table VII. Inflection form of Adverb

Suffix	Stem	Word
-কৈ	ভাল	ভালকৈ
-ই	এনে	এনেই
-এ	গোপন	গোপনে
-আই	বহল	বহলাই
-এ	ডাঙৰ-ডাঙৰ	ডাঙৰে-ডাঙৰে
-তা	গাঢ়	গাঢ়তা
-তে	বেগ	বেগেতে

D. Adjective

Adjective modifies the noun or pronoun. Adjective are classified into- Predicate adjective, demonstrative adjective, Indefinite adjective, Interrogative adjective, Possessive adjective and Attributive adjective. In Assamese Adjectives are inflected with case, gender, number only if it is used as noun. Table VIII shows examples of some infections of adjectives.

Table VIII. Inflection form of Adjectives

Suffix	Stem	Word
-য	আলস	আলস্য
-ঈ	পৰাক্রম	পৰাক্রমী
-ই	আসক্ত	আসক্তি
-তা	অধম	অধমতা
-তৰ	কঠিন	কঠিনতৰ
-বিলাক	সৰু	সৰুবিলাক
-বোৰ	ভাল	ভালবোৰ
-জনী	ওখ	ওখজনী
-টোক	ডাঙৰ	ডাঙৰটোক

IV. PROPOSED METHODOLOGY

To implement this stemmer, we have created a list of 108 suffixes. It is completely based on Assamese script. According to our algorithm it removes the longest possible suffix present in the list. Following Algorithm defines the working of the system. The same is shown using a flowchart in fig 2.

PROPOSED ALGORITHM

Input: Input word or sentence to extract stem
Step1: Tokenize the sentence into words
Step2: Check whether the suffix of the word matches the suffix list.
a. If yes, then strip off the suffix.
b. If no, then it's a word without stem.
Step 3: Display the word.

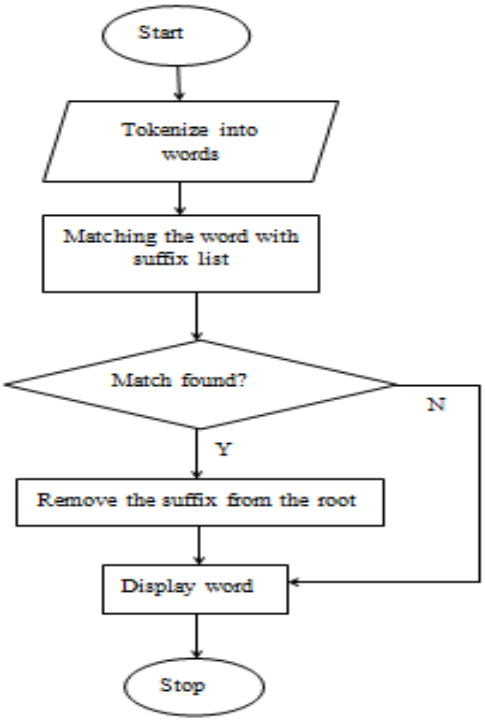

Fig 2. Flowchart of Stemming Algorithm

V. EVALUATION

We developed a system using Python on the basis of the above mentioned algorithm. To evaluate the performance of the system accuracy is calculated on the behalf of 20,000 words. . Also, Accuracy of a system depends on the number of rules that are used for stripping suffix. To measure the accuracy following equation is used.

$$Accuracy\ (\%) = \frac{Correct\ Stem}{Total\ Number\ of\ Given\ Words}\ X\ 100$$

Table IX and X illustrate the statistics of results produced through the above equation.

Table IX. Obtained result for correctly stemmed words

Total word	20000
Correctly Stemmed	18873
Stem as single character inflection	7705
Stem as multiple character inflection	2089
No inflection	9079

Table X. Obtained result for incorrectly stemmed words

Total word	20000
Incorrectly Stemmed	1127
Stem as single character inflection	268
Stem as multiple character inflection	152
No inflection	707

Among 20,000 words, 18873gives correct stem and 1127 words are wrongly stemmed because Assamese has single character suffix more than other languages, such single character are – ‘ক’ , ‘ত’ , ‘ৰ’ etc. cause over-stemming. Sometimes it considers the character from a word which do not require stemming, this is also one of reasons for occurrence of over-stemming. The result of stem was 94.36%. E.g. ‘কৰ’ has no suffix, but as ‘ৰ’ is a single character suffix that’s why ‘ৰ’ is removed from the word which gives us result as ‘ক’. Similarly, ‘জন’ is in the suffix list so it also removes ‘জন’ from প্রয়োজন, উপার্জন, আয়োজন although ‘জন’ should not be removed from these words. Some correctly stemmed words are shown table 11.

Table XII. Example of correctly stemmed words

Word	Stem
ভাৰতীয়	ভাৰত
নাম	নাম
ছোৱালীজনীৰ	ছোৱালী
মানুহজন	মানুহ
বিদ্যার্থী	বিদ্যা
পুৱাৰৈ	পুৱা
কৰক	কৰ
সকলোকে	সকলো
কিতাপখন	কিতাপ
মাস্টৰী	মাস্টৰ
তেজপুৰীয়া	তেজপুৰ

VI. CONCLUSION

In this paper, we have shown an approach of developing a stemmer for Assamese. A rule-based approach is applied for this. We have studied different inflectional words of Assamese and developed 108 rules of stemming. In order to evaluate the system, we tested our approach with 20000 words. Out of these 20000, the system was correctly identifying 94.36% stems. In future, we wish to improve our stemmer. One of the possible approaches can be a ripple down approach.

REFERENCES

1. B. L. Julie, “Development of a Stemming Algorithm”, Mechanical Translation and Computational Linguistics, Vol. 11, pp. 22-31, 1968.
2. M.F. Porter, “An algorithm for suffix stripping”, Program, Vol. 14 Issue: 3, pp.130-137, 1980.
3. R. Ananthakrishnan and R. D. Durgesh, “A Lightweight Stemmer for Hindi”, EACL, 2003.
4. M. M. Mudassar and S. J. Tanveer, “Discovering suffixes: A Case Study for Marathi Language”, International Journal on Computer Science and Engineering, Vol. 02, pp. 2716-2720, 2010.
5. A. T. Eiman and L. Jessica, ” Towards an Error-Free Arabic Stemming”, ACM workshop on Improvingnon english web searching, pp. 9-15, 2008.
6. L. S. Leah, C. E. Margaret and A. Nasreen, ” Hindi CLIR in Thirty Days”, ACM Transactions on Asian Language Information Processing, Vol. 2, pp. 130-142, 2003.
7. A. Juhi, J. Nisheeth and M. Iti, “A Lightweight Stemmer for Gujarati”, Annual Convention of Computer Society of India, 2012.
8. A. Qurat-ul-Ain, N. Asma and H. Sarmad, “Assas-Band, an Affix-Exception-List Based Urdu Stemmer”, Asian Language Resources, pp.40-47, 2009.
9. S. Sandipan and B. Sivaji, “Design of a Rule-based Stemmer for Natural Language Text in Bengali” , IJCNLP-08 workshop on NLP for Less Privileged Languages, pp. 65-72, 2008.
10. P. D. Chandrakant and P. M. Jayesh, “GUJSTER: a Rule based stemmer using Dictionary Approach”, International Conference on Inventive Communation and Computational Technologies, pp. 496-499, 2017.
11. P.D. Chris, “Another Stemmer ”, Association for Computing Machinery, Vol. 24, 1990.
12. J. Marie-Clarie and S. Dan, ” Conservative stemming for search and indexing”, ”, Association for Computing Machinery, 2005.
13. S. Navanath, S. Utpal and K. Jugal,” Analysis and Evaluation of Stemming Algorithms: A case Study with Assamese”, International Conference on Advances in Computing, Communications and Informatics, pp. 842-846, 2012.
14. S. Navanath, S. Utpal and K. Jugal, “A Suffix-based Noun and Verb Classifier for an Inflectional Language”, International Conference on Asian Language Processing, pp. 19-22, 2010.
15. S. Navanath, K. M. Kishori, S. Utpal, K. K. Jugal, “An Improved Stemming Approach Using HMM for a Highly Inflectional Language”, International Conference on Intelligent Text Processing and Computational Linguistics, pp. 164-173, 2013.

16. S. Navanath, S. Utpal and K. Jugal, "Stemming resource-poor Indian languages", ACM Transactions on Asian Language Information Processing, Vol. 13, 2014.
17. S. Padmaja, S. Utpal and K. Jugal," Suffix Stripping Based NER in Assamese for Location Names", National Conference on Computational Intelligence and Signal Processing, pp. 91-94, 2012.

AUTHORS PROFILE

Author-1 Photo

Swagata Seal, is a student at Banasthali Vidyapith. Her areas of interest are natural language processing and machine translation.

Author-2 Photo

Nisheeth Joshi, is an Associate Professor at Bnasthali Vidyapith, India. He primiraliy works in Machine Translation, Information Retrieval, Cognitive Computing. He has over 12 years of teaching experimence.