An Innovative Lemmatization Technique for Bangla Nouns by using Longest Suffix Stripping Methodology in Decreasing Order

Alok Ranjan Pal
Detp. of Computer Sc. and Engg.
College of Engg. & Mgmt, Kolaghat
Kolaghat, India
chhaandasik@gmail.com

Niladri Sekhar Dash
Linguistic Research Unit
Indian Statistical Institute
Kolkata, India
ns_dash@yahoo.com

Diganta Saha
Dept. of Computer Sc. and Engg.
Jadavpur University
Kolkata, India
neruda0101@yahoo.com

***Abstract*-In this proposed work, an attempt is made to find out the root part from inflected Bangla nouns by applying an innovative technique by using longest suffix stripping methodology in decreasing order. The test data is generated from a Bangla text corpus developed in the TDIL Project of the Govt. of India. The exhaustive suffix list obtained from the research work carried out at *Linguistic Research Unit of Indian Statistical Institute, Kolkata* while the Bangla non-inflected noun list used in this work is obtained from a wordlist generated by *Pashchimbanga Bangla Akademi*, Kolkata and available in the net. The algorithm is applied on randomly selected 1273 noun instances and accuracy is achieved around 94%.**

Keywords—lemmatization; suffix stripping; part-of-speech tagging; morphology

I. INTRODUCTION

Nouns in most of the natural languages are available either as single bases in their non-inflected forms or in affixed bases in inflected forms. In case of the second option, nouns are inflected in two ways: inflectional morphology (e.g., cow-cows, baby-babies, ox-oxen, etc.) and derivational morphology (e.g., organization-organizational, computer-computerized, etc.). In most cases the range of inflection is restricted within a few types. In Bangla - a language of the Indian sub-continent, the range of inflection types is quite large. Here a noun can have several types of inflection. For instance, the words *māthā* "head" can have several inflected forms, such as, *māthāte* "in head", *māthā□ā* "the head", *māthāy* "on head", *māthār* "of head", *māthāgulo* "heads", *māthārā* "heads", *māthā□i* "the head", *māthāder* "of heads", *māthāri* "of head itself", etc. This happens in case of single inflection only, which can be doubled if the processes of double inflection addition is invoked, such as, *māthātei* "in head itself", *māthā□āi* "the head itself", *māthāyo* "on the head also", *māthāri* "of head itself", *māthāgulote* "on the heads", *māthārāi* "heads themselves", *māthā□io* "the head also", *māthāderi* "of heads themselves", etc. There are few determiners like, -*□ā,* -*□i, -khānā, -khāni, -khānek* as well as a few emphatic markers, like -*i* and -*o*, etc. which are often tagged with inflected nouns for double inflection. So, finding base forms of Bangla nouns is a challenging task. The situation becomes far more complicated in case of nouns made with double suffixes like *māthāgulote* "in heads" (two suffixes: -*gulo* and -*te*) and multiple suffixes, such as, *māthāgulotei* "in heads themselves" (three suffixes: -*gulo*, -*te* and -*i*).

In the proposed approach, an algorithm has been designed to find the root part from inflected nouns by stripping the longest suffix in reserve order. To increase the efficiency of the algorithm, an alphabetically sorted list of non-inflected words has been used in this work. The exhaustive suffix list obtained from the research work carried out at *Linguistic Research Unit of Indian Statistical Institute, Kolkata* while the Bangla non-inflected noun list used in this work is obtained from a wordlist generated by *Pashchimbanga Bangla Akademi*, Kolkata and available in the net.

The organization of the paper is as follows: Section 2 describes the survey on lemmatisation and stemming in Indian languages; nature of Bangla nominal morphology is discussed in Section 3; Section 4 depicts the overall approach in detail; results and corresponding evaluations are presented in Section 5; Section 6 concludes the discussion.

II. SURVEY ON LEMMATISATION AND STEMMING IN INDIAN LANGUAGES

Due to limited number of morphological variants many works on lemmatization and stemming have been done for English. For Bangla, however, not much work is done due to its morphological complexities involved in word formation. Since the language is very rich in morphological diversity, less amount of work on lemmatisation or stemming has been done in it. This does not stand true for Bangla alone, this is equally true for most of the Indian languages as the following references show:

Ramanathan and Rao [1] designed a lightweight stemmer for Hindi language in 2003. The approach was tested on 35977 unique words taken from different sources like politics, sports, business, health, etc. The error in this method was 4.68% for under-stemming and 13.84% for over-stemming task. Majumder et al. [2] proposes a statistical model for stemming named as YASS (Yet Another Suffix Stripper). It is a clustering-based approach that works on string distance measures sans linguistic knowledge. The model clusters the

978-1-4673-7309-8/15/$31.00 ©2015 IEEE

lexicons generated from a text corpus into some homogenous groups and each group represents an equivalent class containing morphological variants of a single root word. Dasgupta and Nag [3] designs a strategy for unsupervised morphological processing of Bangla words. This method is tested on a set of 4110 human-segmented Bangla words with 83% success rate. Pandey and Siddiqui [4] proposes an unsupervised stemming method for Hindi words based on Goldsmith approach. This method is evaluated on 1000 words randomly selected from the Hindi WordNet and it records 89.9% accuracy. Majgaonker and Siddiqui [5] proposes an unsupervised stemming approach for Marathi words. When the algorithm is tested on 1500 manually stemmed words, it comes out with 82.5% accuracy. Suba et al. [6] develop two kinds of stemmers for Gujarati words. The first one is a light-weight inflectional stemmer based on a hybrid approach, while the second one is a heavy-weight derivational stemmer based on a rule-based approach. While the inflectional stemmer produces around 90.7% accuracy, the derivational stemmer gives around 70.7% accuracy. Zahurul Islam, et al. [7] proposes a light-weight stemmer for Bangla which can be used in a spell checker with some satisfactory outputs. Dash [13] has recently proposed a simple linguistic strategy for lemmatizing Bangla inflected nouns, which may be further converted into a rule-based approach studded with algorithms for stemming for lemmatizing not only inflected nouns but also inflected verbs and adjectives of the language. Along this line some works may be credited to Chaudhuri [10], Sarkar and Bandyopadhyay [11], and Dasgupta and Khan [12]. Most of these works are based on rule-based approach for stemming and lemmatization.

III. Nature of Bangla Nominal Morphology

In Bangla, the scale of morphological variety is so large that no defined rule can cover all types of inflections for nouns. For example, the following instances can be solved by single-suffix stripping approach:

(1) baiguli = bai + (-guli)

(2) baite = bai + (-te).

But, this becomes more challenging when more than one suffix remains appended with the word. For example, the following instances can be solved by multiple-suffix stripping approach:

(3) baigulite = bai + (-guli) + (-te)

(4) baitite = bai + (t-i) + (-te)

And, the following instances are the obstacles in the lemmatization process, as the identification of accurate suffix part is quite complicated:

(5) maayer = maa + (-yer) is right

(6) chabiTi = chabi + (-Ti) is right

But,

(7) samayer = sama + (-yer) is wrong.

(8) maaTi = maa + (-Ti) is wrong.

The proposed approach takes the tagged text as input and generates the lemmatized form of the individual noun.

IV. The Proposed Approach

In the proposed approach, the tagged nouns are given as input and the base forms of the inflected nouns are generated as output.

First, the Bangla text corpus, developed in the TDIL project of the Govt. of India is taken as an input. But, the texts in the corpus are non-normalized in nature. So, these texts are passed through a series of manual text normalization steps (refer section V.*A*).

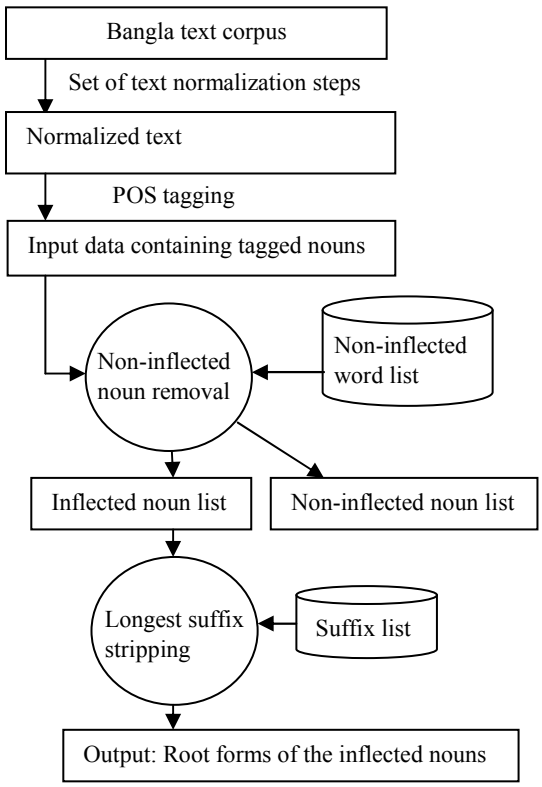
After that, the nouns in the normalized texts are tagged (refer section V.*D*).

Next, the non-inflected nouns are detected and removed from the test data by the help of the non-inflected word list developed by the *Bangla Academy*. This procedure was applied to increase the accuracy of the result, as the last characters of several non-inflected nous, match with different suffixes.

Then, the inflected nouns are passed through the longest-match suffix stripping algorithm. The suffix list is collected from the *Linguistic Research Unit of Indian Statistical Institute, Kolkata* [13]. To match the longest suffix with the end part of the inflected nouns, the suffixes are rearranged according to their lengths in decreasing order (refer section V.*C*).

Thus the suffix stripped root forms of the inflected nouns are obtained as output. The overall approach is represented by the following diagram (refer Fig. 1).

Fig.1. Pictorial representation of the overall approach



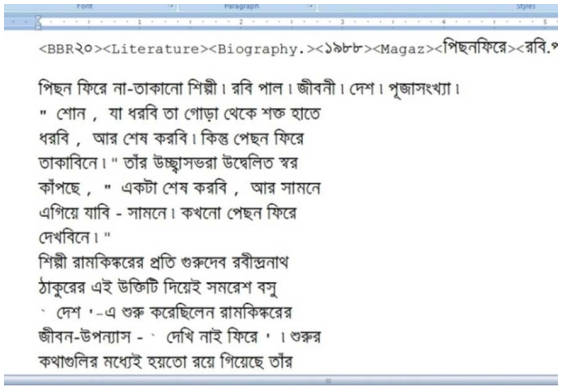
V. Results and Corresponding Evaluations

The following steps were executed to evaluate the system-

A. Non-normalized texts

The text, retrieved from the Bangla text corpus was non-normalized in nature (refer Fig. 2).

Fig.2. A sample non-normalized text from the Bengali corpus



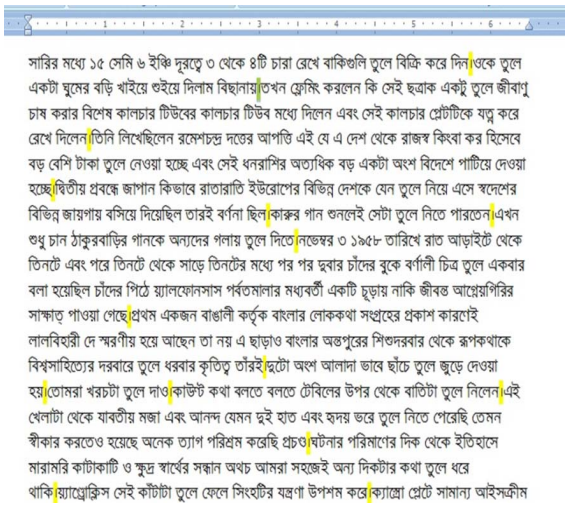
<BBR২০><Literature><Biography.><১৯৮৮><Magaz><পিছনফিরে><রবি.৭

পিছন ফিরে না-তাকানো শিল্পী। রবি পাল। জীবনী। দেশ। পূজাসংখ্যা।
" শোন , যা ধরবি তা গোড়া থেকে শক্ত হাতে
ধরবি , আর শেষ করবি। কিন্তু পেছন ফিরে
তাকাবিনে।" তাঁর উচ্ছ্বাসভরা উদ্বেলিত স্বর
কাঁপছে , " একটা শেষ করবি , আর সামনে
এগিয়ে যাবি - সামনে। কখনো পেছন ফিরে
দেখবিনে।"
শিল্পী রামকিঙ্করের প্রতি গুরুদেব রবীন্দ্রনাথ
ঠাকুরের এই উক্তিটি দিয়েই সমরেশ বসু
` দেশ '-এ শুরু করেছিলেন রামকিঙ্করের
জীবন-উপন্যাস - ` দেখি নাই ফিরে ' । শুরুর
কথাগুলির মধ্যেই হয়তো রয়ে গিয়েছে তাঁর

B. Normalized text

The non-normalized text was passed through a series of manual normalization steps, as a) fonts were similarized, b) un-even newlines, blanks are removed, c) comma, colon, semi-colon, hyphen, under score, slashes etc were removed, d) special Bangla sentence termination symbol "।" (dāri) was taken into account etc. and the text was generated as below (refer Fig. 3)-

Fig. 3. A sample normalized text



সারির মধ্যে ১৫ সেমি ৬ ইঞ্চি দূরত্বে ৩ থেকে ৪টি চারা রেখে বাকিগুলি তুলে বিক্রি করে দিন।ওকে তুলে একটা ঘুমের বড়ি খাইয়ে শুইয়ে দিলাম বিছানায়।তখন ফ্রেমিং করলেন কি সেই ছত্রাক একটু তুলে জীবাণু চাষ করার বিশেষ কালচার টিউবের কালচার টিউব মধ্যে দিলেন এবং সেই কালচার প্লেটটিকে যত্ন করে রেখে দিলেন।তিনি লিখেছিলেন রমেশচন্দ্র দত্তের আপত্তি এই যে এ দেশ থেকে রাজস্ব কিংবা কর হিসেবে বড় বেশি টাকা তুলে নেওয়া হচ্ছে এবং সেই ধনরাশির অত্যধিক বড় একটা অংশ বিদেশে পাটিয়ে দেওয়া হচ্ছে।দ্বিতীয় প্রবন্ধে জাপান কিভাবে রাতারাতি ইউরোপের বিভিন্ন দেশকে যেন তুলে নিয়ে এসে স্বদেশের বিভিন্ন জায়গায় বসিয়ে দিয়েছিল তারই বর্ণনা ছিল।কারুর গান শুনলেই সেটা তুলে নিতে পারতেন।এখন শুধু চান ঠাকুরবাড়ির গানকে অন্যদের গলায় তুলে দিতে।নভেম্বর ৩ ১৯৫৮ তারিখে রাত আড়াইটে থেকে তিনটে এবং পরে তিনটে থেকে সাড়ে তিনটের মধ্যে পর পর দুবার চাঁদের বুকে বর্ণালী চিত্র তুলে একবার বলা হয়েছিল চাঁদের পিঠে য়্যালফোনসাস পর্বতমালার মধ্যবর্তী একটি চূড়ায় নাকি জীবন্ত আগ্নেয়গিরির সাক্ষাত্ পাওয়া গেছে।প্রথম একজন বাঙালী কর্তৃক বাংলার লোককথা সংগ্রহের প্রকাশ কারণেই লালবিহারী দে স্মরণীয় হয়ে আছেন তা নয় এ ছাড়াও বাংলার অন্তপুরের শিশুদরবার থেকে রূপকথাকে বিশ্বসাহিত্যের দরবারে তুলে ধরবার কৃতিত্ব তাঁরই।দুটো অংশ আলাদা ভাবে ছাঁচে তুলে জুড়ে দেওয়া হয়।তোমরা খরচটা তুলে দাও।কাউন্ট কথা বলতে বলতে টেবিলের উপর থেকে বাতিটা তুলে নিলেন।এই খেলাটা থেকে যাবতীয় মজা এবং আনন্দ যেমন দুই হাত এবং হৃদয় ভরে তুলে নিতে পেরেছি তেমন স্বীকার করতেও হয়েছে অনেক ত্যাগ পরিশ্রম করেছি প্রচুর।ঘটনার পরিমাণের দিক থেকে ইতিহাসে মারামরি কাটাকাটি ও ক্ষুদ্র স্বার্থের সন্ধান অথচ আমরা সহজেই অন্য দিকটার কথা তুলে ধরে থাকি।য়্যাগ্রেক্লিস সেই কাঁটাটা তুলে ফেলে সিংহটির যন্ত্রণা উপশম করে।ক্যান্ত্রো প্লেটে সামান্য আইসক্রীম

C. Suffix list

The suffix list, used in this work was developed recently by a team of scholars at Linguistic Research Unit, Indian Statistical Institute, Kolkata and this has been tested by some experts who have checked the list with a range of inflected nouns taken from the Bangla corpus and validated the list (Dash 2014) [13]. For implementation perspective, the suffixes were rearranged in decreasing order based on their length (refer Fig. 4). According to the morphological feature of the Bengali nouns, they end with no-suffix or single suffix or multiple suffixes. This algorithm strips the suffix of longest match from a noun to find the root form of it.

Fig. 4. The suffixes in decreasing length

দিগেরেও,খানাতেই,খানাতেও,খানিকেই,খানিকেও,খানিতেই,খানিতেও,টুকুকেই,টুকুকেও,টুকুতেই,টুকুতেও,গুলি কেই,গুলিকেও,গুলিতেই,গুলিতেও,গুলোকেই,গুলোকেও,গুলোতেই,গুলোতেও,গুলাকেই,গুলাকেও,গুলাতেই,গুলা তেও,দিগেরেও,খানাকে,খানাতে,খানিকে,খানিতে,টুকুকে,টুকুতে,গুলিকে,গুলিতে,গুলোকে,গুলোতে,গুলাকে,গুলা তে,দিগেরে,টাকেরই,টাকেরও,খানারও,খানায়ও,খানিরই,খানিরও,টুকুরই,টুকুরও,গুলিরই,গুলিরও,গুলোরই,গু লোরও,গুলোয়ই,গুলোয়ও,গুলারই,গুলারও,গুলায়ও,দিগেরও,দিগেরই,দিগেরই,খানারই,খানায়ই,গুলায়ই,দেরকে ই,দেরকেও,দিগকেই,দিগকেও,খানাই,খানাও,খানিই,খানিও,টুকুই,টুকুও,গুলিই,গুলিও,গুলোই,গুলোও,গুলাই,গু লাও,খানার,খানায়,খানির,টুকুর,গুলির,গুলোর,গুলোয়,গুলার,গুলায়,দেরকে,দিগকে,দিগের,টাকেই,টাকেও,টাতে ই,টাতেও,টিকেই,টিকেও,টিতেই,টিতেও,খানা,খানি,টুকু,গুলো,গুলি,গুলা।েরাই।েরাও,দেরই,দেরও,দিগই,দিগ ও,টাকে,টাতে,টিকে,টিতে।েতেই।েতেও,টারই,টারও,টায়ই,টায়ও,টিরই,টিরও,য়েরই,য়েরও।েরা,দের,দিগ।ে তে,য়ের,টিই,টিও,টাই,টাও,রাই,রাও,টার,টায়,টির,কেই,কেও,তেই,তেও।েরই।েরও।ায়ই।ায়ও।ারই।ারও ।ায়ই।ায়ও,টা,টি,রা।ের,কে,তে,কে।ার।ায়।েই।েও,রই,রও,ই,ও,র।ে

D. Partial view of the input data

The nouns in the input data were tagged in the following way. The input text contained total 1273 tagged nouns (refer Fig. 5).

Fig. 5. The tagged input data

আধুনিক হিসাবশাস্ত্র/noun ক্রমশঃই জটিল হইতে জটিলতর আকার/noun ধারণ/noun
করিতেছে,শতাব্দীর/noun পর শতাব্দী/noun ধরিয়া ব্যবসায়িগণ ও রাষ্ট্রের/noun রাজস্ববিভাগ কর্তৃক
ব্যবহারিক ক্ষেত্রে হিসাবের/noun সর্বাধিক প্রয়োগের/noun ফলে বর্তমানে হিসাবশাস্ত্র ব্যবহারিক
বিজ্ঞানে/noun মর্যাদা/noun লাভ করিয়াছে,হিসাবশাস্ত্রের/noun ক্রমবিকাশের/noun পিছনে রাষ্ট্র/noun
ও ব্যবসায়িগণের/noun ভূমিকা/noun অনস্বীকার্য,পরিবর্তিত পরিস্থিতিতে/noun অর্থাত্ সময়ের/noun
পরিবর্তনের/noun সঙ্গে সঙ্গে মূলতঃ ইহাদের ব্যবহারিক সুবিধা/noun এবং উপযোগের/noun দিকে
লক্ষ্য/noun রাখিয়াই নূতন নূতন নীতি/noun হিসাবশাস্ত্রে/noun সংযোজিত হইতেছে।এই সকল নীতি ও
সূত্রগুলি/noun শিক্ষকগণ এবং শিক্ষাপ্রাপ্ত ব্যক্তিগণের অভিজ্ঞতালব্ধ মূল্যবান
ফসল/noun।হিসাবশাস্ত্র/noun সম্পর্কে পূর্ণাঙ্গ জ্ঞান/noun লাভ করিতে হইলে ইহার ক্রমবিকাশের/noun
ইতিহাস ইহার প্রকৃতি কার্যাবলী/noun ইত্যাদি সম্পর্কে পরিষ্কার ধারণা/noun থাকা
প্রয়োজন/noun।সেইজন্য এই অধ্যায়ে/noun উক্ত বিষয়াদি সম্পর্কে বিস্তারিত/noun আলোচনা/noun
করা হইল।হিসাবশাস্ত্রের/noun ভিত্তি মানবসভ্যতার/noun ক্রমবিবর্তনের/noun ইতিহাসের/noun মতই
হিসাবশাস্ত্রের/noun ক্রমবিবর্তনের/noun ইতিহাস/noun অতি প্রাচীন ও বৈচিত্রময়,প্রাচীন যুগ/noun

E. Partial view of the output file

The partial view of the output file is given below (refer Fig. 6).

Fig. 6. The generated output data

আধুনিক হিসাবশাস্ত্র/NOUN/হিসাবশাস্ত্র ক্রমশঃই জটিল হইতে জটিলতর আকার/NOUN/আকার ধারণ/NOUN/ধারণ করিতেছে ।শতাব্দীর/NOUN/শতাব্দী পর শতাব্দী/NOUN/শতাব্দী ধরিয়া ব্যবসায়িগণ ও রাষ্ট্রের/NOUN/রাষ্ট্র রাজস্ববিভাগ কর্তৃক ব্যবহারিক ক্ষেত্রে হিসাবের/NOUN/হিসাব সর্বাধিক প্রয়োগের/NOUN/প্রয়োগ ফলে বর্তমানে হিসাবশাস্ত্র ব্যবহারিক বিজ্ঞানে/NOUN/বিজ্ঞান মর্যাদা/NOUN/মর্যাদা লাভ করিয়াছে ।হিসাবশাস্ত্রের/NOUN/হিসাবশাস্ত্র ক্রমবিকাশের/NOUN/ক্রমবিকাশ পিছনে রাষ্ট্র/NOUN/রাষ্ট্র ও ব্যবসায়িগণের/NOUN/ব্যবসায়িগণ ভূমিকা/NOUN/ভূমিকা অনস্বীকার্য ।পরিবর্তিত পরিস্থিতিতে/NOUN/পরিস্থিতি অর্থাত্ সময়ের/NOUN/সম পরিবর্তনের/NOUN/পরিবর্তন সঙ্গে সঙ্গে মূলতঃ ইহাদের ব্যবহারিক সুবিধা/NOUN/সুবিধা এবং উপযোগের/NOUN/উপযোগ দিকে লক্ষ্য/NOUN/লক্ষ্য রাখিয়াই নূতন নূতন নীতি/NOUN/নীতি হিসাবশাস্ত্রে/NOUN/হিসাবশাস্ত্র সংযোজিত হইতেছে ।এই সকল নীতি ও সূত্রগুলি/NOUN/সূত্র শিক্ষকগণ এবং শিক্ষাপ্রাপ্ত ব্যক্তিগণের অভিজ্ঞতালব্ধ মূল্যবান ফসল/NOUN/ফসল ।হিসাবশাস্ত্র/NOUN/হিসাবশাস্ত্র সম্পর্কে পূর্ণাঙ্গ জ্ঞান/NOUN/জ্ঞান লাভ করিতে হইলে ইহার ক্রমবিকাশের/NOUN/ক্রমবিকাশ ইতিহাস ইহার প্রকৃতি কার্যাবলী/NOUN/কার্যাবলী ইত্যাদি সম্পর্কে/NOUN/সম্পর্কে পরিষ্কার ধারণা/NOUN/ধারণা থাকা প্রয়োজন/NOUN/প্রয়োজন ।সেইজন্য এই অধ্যায়ে/NOUN/অধ্যায় উক্ত বিষয়াদি সম্পর্কে বিস্তারিত/NOUN/বিস্তারিত আলোচনা/NOUN/আলোচনা করা হইল ।হিসাবশাস্ত্রের/NOUN/হিসাবশাস্ত্র ভিত্তি মানবসভ্যতার/NOUN/মানবসভ্যত

VI. Result and Discussion

The model was evaluated on 1273 noun instances and 94% accuracy was achieved.

The efficiency of the algorithm is measured based on the three parameters "Precision", "Recall", and "F-Measure" as well. Precision (P) is the ratio of the "matched instances according to the human decision" and "number of instances responded by the system". Recall value (R) is the ratio of "number of instances matched with the human decision" and "total number of instances in the dataset" and F-Measure is evaluated as "(2*P*R / (P+R))".

As the system responded to all the target instances present in the dataset (either correctly or wrongly), the Precision and the Recall values are same. P=R=1197/1273=0.94 and F-Measure= 1.7672/1.88=0.94.

According to the morphological feature of the Bengali nouns, they end with no-suffix or single suffix or multiple suffixes. As the algorithm stripped the longest suffix from the nouns, the root of the nouns were derived in most of the cases, except a few, which are discussed in section VI *B*.

Generally, the Bengali nouns are not deformed with a prefix part, like western languages. So, prefix stripping is not considered for Bengali noun lemmatization.

A. Complexity of the algorithm

The complexity of the algorithm is $O(n^2)$, as for every noun in the test set, the whole list of the suffixes is traversed individually.

B. Few close observations

A few obstacles in this approach have been discussed in Section 3. According to the nature of the Bangla nouns, few words end with the characters which are used as suffix as well. For example (refer Fig. 7)

Fig. 7. Few obstacles in Bengali morphology

হিসাবশার/NOUN/হিসাবশাস্ত্র	কার্য্যব্যবস্থার/NOUN/কার্য্যব্যবসথ	মাথার/NOUN/মাথ
সময়ের/NOUN/সম	ব্যবস্থার/NOUN/ব্যবস্থ	কপালে/NOUN/কপালে
মানবসভ্যতার/NOUN/মানবসভ্যত	ব্যবসায়ে/NOUN/ব্যবসায়	রচনার/NOUN/রচন
গুহার/NOUN/গুহ	পরিচালনায়/NOUN/পরিচালন	লেখায়/NOUN/লেখ
সূচনার/NOUN/সূচন	পরিচালনার/NOUN/পরিচালন	এলাকায়/NOUN/এলাক
প্রস্তরও/NOUN/প্রস্ত	কারখানায়/NOUN/কার	পালনের/NOUN/পালন
সভ্যতার/NOUN/সভ্যত	ব্যয়ের/NOUN/ব্য	হাতে/NOUN/হা
আত্মরক্ষার/NOUN/আত্মরক্ষ	গলার/NOUN/গল	বৃহস্পতিবার/NOUN/বৃহস্পতিব
জায়গায়/NOUN/জায়গ	সঙ্গেরে/NOUN/সঙ্গেরে	চেষ্টাতেই/NOUN/চেষ্ট
জীবিকার/NOUN/জীবিক	বাংলার/NOUN/বাংল	
চেতনার/NOUN/চেতন	কারার/NOUN/কার	
সমস্যার/NOUN/সমস্য	পক্ষার/NOUN/পক্ষ	
বিনিময়ের/NOUN/বিনিম	বুকে/NOUN/বু	

VII. CONCLUSION AND FUTURE WORK

According to the structure of the Bengali nouns, the degradation in accuracy of the performance was obvious. We wish to deal with these erroneous situations by forming new learning rules.

Acknowledgement

The authors acknowledge Mr. Saptarshi Datta, Miss. Trisha Saha and Miss. Sudipta Bose for their spontaneous help to this work.

Reference

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

[2] P. Majumder, M. Mitra, S.K. Parui, G. Kole, P. Mitra, and K. Datta, "YASS: Yet Another Suffix Stripper, Association for Computing Machinery Transactions on Information Systems," 25(4), 2007, pp.18-38.

[3] S. Dasgupta and V. Nag, "Unsupervised Morphological Parsing of Bengali. Language Resources and Evaluation," 40(3-4), 2006, pp. 311-330.

[4] A.K. Pandey and T.J. Siddiqui, "An Unsupervised Hindi Stemmer with Heuristic Improvements. Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data," 2008, pp. 99-105.

[5] M.M. Majgaonker and T.J Siddiqui, "Discovering Suffixes: A Case Study for Marathi Language," International Journal on Computer Science and Engineering, 2(8), 2010, pp. 2716-2720.

[6] K. Suba, D. Jiandani and P. Bhattacharyya, "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati," Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, Chiang Mai, Thailand, 2011, pp. 1-8.

[7] M. Z. Islam, M.N. Uddin and M. Khan, "A Light Weight Stemmer for Bengali and its Use in Spelling Checker," Proceedings of the 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA07), Irbid, Jordan, March, 2007, pp. 19-23.

[8] M.F. Porter, "An algorithm for suffix stripping. Program," 14(3), 1980, pp. 130-137.

[9] P. Kundu and B.B. Chaudhuri, "Error Pattern in Bengali Text," International Journal of Dravidian Linguistics, 1999.

[10] B.B. Chaudhuri, "Reversed word dictionary and phonetically similar word grouping based spell-checker to Bengali text," Proceedings of LESAL Workshop, 2001.

[11] S. Sarkar and S. Bandyopadhyay, "Study on Rule-Based Stemming Patterns and Issues in a Bengali Short Story-Based Corpus," ICON-2009 (Poster).

[12] S. Dasgupta and M. Khan. Morphological parsing of Bangla words using PCKIMMO. In: ICCIT 2004. (2004)

[13] N.S. Dash, "Back to Basics: A Road to Return to Nominal Base through Lemmatization," Proceedings of Abstracts of the 36th International Conference of the Linguistic Society of India (ICOLSI-36), 1-4 December 2014, Dept. of Linguistics, University of Kerala, Trivandrum, Kerala, India. Pp. 94-104.