A rule based bengali stemmer

Conference Paper · September 2014

DOI: 10.1109/ICACCI.2014.6968484

5 authors, including:

Md. Redowan Mahmud
RMIT University
25 PUBLICATIONS **1,169** CITATIONS
SEE PROFILE

Mahbuba Afrin
Swinburne University of Technology
10 PUBLICATIONS **147** CITATIONS
SEE PROFILE

Md. Abdur Razzaque
University of Dhaka
119 PUBLICATIONS **1,410** CITATIONS
SEE PROFILE

A Rule Based Bengali Stemmer

Md. Redowan Mahmud, Mahbuba Afrin, Md. Abdur Razzaque
Green Networking Research Group, Dept. of Computer Science and Eng., University of Dhaka, Bangladesh
Email: ratul06oct, afrinritu92 AT gmail DOT com, razzaque AT cse DOT univdhaka DOT edu

Ellis Miller, Joel Iwashige
Code Crafters International Bangladesh
S.S. Steel Building, Suite 302
17/C Panthapath, Dhaka 1207
Email: Ellis DOT Miller, Joel DOT Iwashige AT CodeCraftersIntl DOT com

***Abstract*—One of the biggest challenges in doing word lookups is to derive the appropriate base word for any given word in Bengali. The basic concept to the solution of the problem is to eliminate inflections from a given word to derive its stem word. Stemmers attempt to reduce a word to its root form using stemming process, which reduces an inflected or derived word to its stem or root form. Existing works in the literature use lookup tables either for stem words or suffixes, increasing the overheads in terms of memory and time. This paper develops a rule-based algorithm that eliminates inflections stepwise without continuously searching for the desired root in the dictionary. To the best of our knowledge, this paper first investigates that, in Bengali morphology, for a large set of inflections, the stems can be computed algorithmically cutting down the inflections step by step. The proposed algorithm is independent of inflected word lengths and our evaluation shows around 88% accuracy.**

Keywords—Bengali; Stemmer; Rule based Stemming; Inflections; Verb-root; Stem word

I. INTRODUCTION

Stemming is the process of extracting stem or root word for a given inflected word. The basic concept of stemming is to reduce different grammatical / word forms to its root, stem or base form**.** Stemming is widely used in Information Retrieval System [1].

Finding stem word is often required by researchers, readers and more specifically, foreigners willing to learn Bengali as their second language for communication and extracting information from any documents, newspapers, etc.

In the literature, different types of stemming approaches including affix stripping, co-occurrence computation, dictionary look-up, longest suffix matching, probabilistic including natural language processing approaches have been proposed. Most approaches are first developed for English, and later adapted for other languages. However, none of these approaches do not work properly for highly inflectional Indo-Aryan (Hindi, Bengali, Marathi, and Gujarati) languages [2].

It is quite difficult to determine the stem words from inflected words in Bengali as it is one of the most morphologically rich languages and it has lots of inflectional and derivational variant forms of a word [3] [4], e.g., root word মানুষ (Man) has inflected or derivational forms মানুষগুলো (People), মানুষেরা(Men), মানুষদেরকে(People), মানুষজন (People) etc. Similarly, root word of the verb করা (Do) has the variants করেছি (Have done), করেছিলাম (Did), করছি (Doing), করবো (Will Do), etc.

Taking cognizance of the matter, we found that basically two parts of speech, Noun and Verb have a wide list of inflectional suffixes in Bengali language. A few adjectives can be inflected as well. Here, the main challenge is to identify and formulate rules in Bengali to determine stem words for a given list of inflected words with an acceptable accuracy.

The stemming problem for Bengali words has been addressed in a number of research works in the literature [3][6][7]. In [8], to determine inflections a tokenized word is checked recursively with a predefined inflection set while others [3][4] eliminate suffixes based on hash table matching. In all cases, excessive database searching is performed. These solutions have substantial time and space complexities and thus they become unsuitable for many applications. In this paper, we have solved this problem of extracting Bengali stem words without using any databases and to the best of our knowledge, this is the first work of its kind. The key contributions of this paper can be summarized as follows:

- Identification of the occurrence of different inflections and their pattern classifications
- Development of a rule-driven stemming algorithm for Bengali words that does not require help of any databases for suffix stripping.
- The proposed algorithm is capable of extracting stem words both from inflected verbs and nouns of any length.
- Our evaluation results state that the proposed stemming algorithm can achieve around 88% accuracy.

The rest of the paper is organized as follows: The Section II gives a brief overview of existing stemmers in literature. How inflections occur in Bengali language is discussed in Section III and the Section IV discusses about our proposed stemming

978-1-4799-3080-7/14/$31.00

procedure and algorithm. The performance of the proposed algorithm is analysed in Section V and the paper is concluded in Section VI.

II. Related Works

For non Indo-Aryan languages, such as English, Lovins stemmer was published in 1968. This single pass and context sensitive stemmer maintains a list of most frequent suffixes and removes the longest suffix[1]. Later Porters stemming algorithm was introduced based on some rule based conditions and the rule can be represented as:<condition><suffix> → <new suffix>[9].

In 2003, Farsi/Persian language stemmer Kazem Taghva, Russell Beckley used Deterministic Finite Automata (DFA) which is a suffix stripping approach. Whereas for Arabic language, Haidar Harmanani and Walid Keirouz proposed a stem based method on rule engine in 2006. For French language, Jacques Savoy uses derivational suffix stripping proposing suffix stripping method in 2006. And for Indo-Aryan language, Anantha Krishnan Ramanathan and Durgesh D Rao, present a rule based lightweight stemmer for Hindi, in 2003. In 2012, Upendra Mishra, Chandra Prakash used suffix stripping approach based on brute force technique for another Hindi Stemmer. Kartik Suba, Dipti Jiandani and Pushpak Bhattacharyya, present a lightweight inflectional stemmer based on hybrid approach and a heavyweight derivational stemmer based on a rule-based approach for Gujarati language in 2011. Navanath Saharia, Utpal Sharma and Jugal Kalita, adopted suffix stripping approach along with a rule engine for Assamese language in 2012[1].

In [4], using a predefined suffix list, stripping the suffixes on a longest match, a lightweight stemmer for Bengali is proposed. In [8], another Bengali stemmer is proposed where the concept of orthographic syllable of Bengali is introduced. In [3], an idea of suffix stripping is proposed using predefined suffix lookup table.

Our proposed stemming algorithm doesn't use any inflectional hash table. There is no overhead on searching for inflections. It checks all possible inflections on runtime and eliminates them in multiple steps, if necessary. Therefore, any wrong elimination can easily be recovered.

III. Inflections In Bengali

According to some linguistic rules, words in any natural language may become inflected [10]. Bengali words are mostly inflected due to verbal and nominal inflections. Some pronominal and a few of adjective inflections are also seen. In this paper, we deal with verb and noun inflections only.

A. Verbal Inflections

A verb consists of two parts, i.e., verb = verb-root + verbs-ending. e.g., কর [kor] + এ [e] = করে [kore] Here, করে [kore] is verb, কর [kor] is the verb-root and এ [e] is the verb-ending.

Bengali verbs are either finite or non-finite. For finite verbs, the verbs ending vary from tense (present, past, future), person (first, second, third), honor (intimate, familiar, formal) perspective, e.g.,

- আমি যাই (I go) and
- আপনারা যাবেন (You will go).

Here, যাই (Go), যাবেন (Will Go) both are inflected form of the root word যাওয়া (Go). The basic forms of verbal inflections due to variation of verbs ending are given below [11]:

TABLE I. VERBAL INFLECTIONS

Tense	1st&2nd Person	2nd Person (Formal & Informal)	1st person	Formally(honor)	Informally(intimate)
Present Indefinite	ই[I]	এন[en]	ইস[is]	এন [en]	এ[e]
Present Continuous	ছ[ch]	ছ, ছেন[che, chen]	ছিস[chis]	ছেন [chen]	ছে[che]
Present Perfect	এছি[echi]	এছ, এছেন[echo,echen]	এছিস[echis]	এছেন [echen]	এছে[eche]
Present Perfect Continous	__	এন[en]		উন [un]	উক[uk]
Past Indefinite	লাম[lam]	লে[le], লেন[len]	লি[li]	লেন [len]	লা [la](লো)[lo]
Past Continous	ছিলাম[chilam]	ছিলে[chile], ছিলেন[chilen]	ছিলি[chili]	ছিলেন [chilen]	ছিল[echilo]
Past Perfect	এছিলা[echilam]	এছিলে[echile],এছিলেন[echilen]	এছিলি[echili]	এছিলেন [echilen]	এছিল[echilo]
Habituatal Past	তাম[tam]	তে[te], তেন[ten]	তিস[tis]	তেন [ten]	ত [ta](তো)[to]
Habituatal Future	ব(বো) [ba](bo)	বে[be], বেন[ben]	ব[ba]	বেন [ben]	বে[be]
Future Continous	তেথাকব[tethakbo]	তেথাকবেন[te thakben]	তেথাকবি[te thakbi]	তেথাকবেন [te thakben]	তেথাকবে[te thakbe]
Future Perfect	এথাকল[ethaklo]	থাকবে[thakbe]	এথাকবি[e thakbi]	এথাকবেন[e thakben]	এথাকবে[e thakbe]
FuturePerfectContinous	___	বেনওএন[ben o en]	তিস[tis]	বেন[ben]	বে[be]

The root part of a verb is called verb-root. The verb-root is the indivisible part of a verb which represents the inherent essence of a word. In Bengali language, there exists almost 1500 or more verb-roots that can be categorized in 20 different types [11]:

TABLE II. CATEGORIES OF VERB-ROOT

No	Category	Example	No	Category	Example
01	হ[ha]	ক্ষ[khk] ,হ [ha], ল[la] only 3	11	লাফা[lafa]	কাটা[kata], ডাকা[daka], বাজা[baja], আগা[aga] etc. almost 200
02	খা[Kha]	খা[kha] ,ধা[Dha], পা [pa], যা [ja] only 4	12	নাহা[naha]	গাহা[gaha] etc.
03	দি[Di]	দি [di], নি[ni]only 2	13	ফিরা[phira]	ছিটা[chita], শিখা[shikha], ঝিমা[jhima], চিরা[cira] etc. almost 40
04	শু[shu]	শু[shu], ধু [dhu], নু [nu] etc. only 8	14	ঘুরা[ghura]	উচাঁ[uca], লুকা[luka], কুড়া[kura] etc. almost 53
05	কর্[kor]	কর[kor], কম[kom], গড় [gor], চল[col] etc. almost 100	15	ধোয়া[dhoa]	শোয়া[shoa],খোঁচা[khoca],খোয়া[khoa],গোছা[gocha], যোগা[joga] etc. almost 27
06	কহ্[koh]	কহ[koh], সহ[soh], বহ[boh] etc.	16	দৌড়া[doura]	পোঁছা[poucha], দৌড়া[doura] etc.
07	কাট্[kat]	গাঁথ্[gath], চাল্[cal], আঁক্[ak], বাঁধ[badh], কাঁদ[kad] etc. almost 128	17	চটকা[cotka]	সমঝা[somjha],ধমকা[dhomka],কচলা[kocla] etc. almost 100
08	গাহ্[Gah]	চাহ্[cah], বাহ্[bah], নাহ্ [nah]etc.	18	বিগড়া[bigra]	হিচঁড়া[hicra], ছিটকা[chitka], সিটকা[sitka] etc. almost 12
09	লিখ্[likh]	কিন্[kin], ঘির্[ghir], জিত্[jit], ফির্[fir], ভির্[vir], চিন্ [cin] etc. almost 28	19	উলটা[ulta]	দুমড়া[dumra],মুচড়া[mucra], উপচা[upca] etc. almost 27
10	উঠ্[uth]	উড়্[ur], শুন্[shun], ফুট্[phut], খুঁজ্[khuj], খুল্[khul], ডুব্[dub], তুল্[tul] etc. almost 80	20	ছোবলা [chobla]	কোঁচকা[kocka],কোঁকড়া[kokra],কোদরা[kodra] etc.

Here, the category numbers 01-04 are of one length, that means the verb-root consists of only one letter. The category numbers 05-16 are of two length. And, the category numbers 17-20 are of three. That means no verb-root is found in Bengali having more than three. From a wide observation over the Bengali stem words, we notice that:

- Almost all stem words are of length not more than 3 (without diacritic mark)
- Except the single length stem (দে [de], খা [kha]), others are ended with either "া" or no diacritic marked letter.
- Often stem words have no dictionary references. Dictionary uses some widely used versions of that stem word and they are ended with "া" or "ওয়া" [oa] (for single length stem) [11] e.g. দে [de] >দেওয়া [deoa] (Give), নে [ne] >নেওয়া [neoa] (Take), খা [kha] >খাওয়া [khaoa] (Eat), কর [kor] >করা[kora] [Do], ধর [dhor] >ধরা [dhora] (Catch), etc.
- Due to স্বরসঙ্গতি (chord) and inflections of the given word, the first diacritic mark of stem is often changed. e.g. বুঝানো[bujhano] >বোঝা [bojha] (Understand), পেয়েছেন [peyechen] >পাওয়া [paoa] (Get),লিখেছিলাম[likhechilam] >লেখা [lekha] (Write), etc.

B. Noun Inflection

In Bengali, noun inflections occur due to different cases like nominative, objective, genitive and locative. These cases also differ for singular and plural. Usually, singular noun inflections are formed by the nouns ending with রা [ra], টা[ta] টি [ti], খানা[khana], etc. and plural noun inlfections are formed by the nouns ending with এরা [era], গুলি[guli], গুলো[gulo] etc. [12]. Table III shows a number of noun inflections.

TABLE III. NOUN INFLECTION

Case	Animate		Inanimate	
	Singular	***Plural***	***Singular***	***Plural***
Nominative	ছেলেটা (The boy)	ছেলেরা (The boys)	ছাতাটা (The Umbrella)	ছাতাগুলো (The Umbrellas)
Objective	ছেলেটাকে (The boy)	ছেলেদেরকে (The boys)	ছাতাটা (The Umbrella)	ছাতাগুলো (Umbrellas)
Genitive	ছেলেটার (The boy's)	ছেলেদের (The boys')	ছাতাটার(The Umbrella's)	ছাতাগুলোর (The Umbrellas')
Locative			ছাতাটাতে (The Umbrella)	ছাতাগুলোতে (The Umbrellas)

We observe the following facts for noun inflections:

- Noun inflections are limited
- Unlike verbal inflections, they never modify the corresponding stem words

IV. PROPOSED STEMMING ALGORITHM

In this section, we present two separate stemming algorithms- one for the verbal inflected words and another for noun inflected words.

We use hierarchical approach for stripping suffixes from the inflected words. It is based on the idea that a set of commonly used suffixes in the Bengali language can be divided into a combination of smaller and simpler suffixes. Identifying and removing these suffixes stepwise will leave the stem. This differs from traditional Porter algorithm by implicitly maintaining a suffix list in the algorithm itself. This assumption doesn't need to follow any stem dictionary or suffix look up table. This algorithm not only eliminates the inflections, but also illuminates different variations of stems in Bengali language described in the sections III(A) and III(B).

A. Stemming Algorithm for Verbal Inflection

In developing this algorithm, the basic task is to identify and categorize the inflections and finding some patterns among them. From Table 1, we categorize verbal inflections in two types:

- Independent inflections
- Combined inflections

Independent inflections are those minimum length (1 or 2) inflectional suffixes that are used independently and exclusively in a word. Only one step inflection elimination is sufficient to retrieve stem from these kinds of inflected words. The two possible set of Independent inflections are

- Set-1: {ই[i], ছ[ch], ব[b], ল[l], ত[t], ন[n], ক[k], ম[m], স[s]}, e.g., করি [kori], বলছ [bolcho], চলব [colbo] , করল [korlo], লিখত [likhto], চলন [colon], etc.

- Set-2:{লা[la], লো[lo], তো[to], লে[le], তা[ta], তি[ti], ছি[chi], ছে[che], ছো[cho], তে[te], লি[li], বে[be]}, e.g., করলা [korla], বললে [bolle], খেলতা[khelto], করছি [korchi], চলছে [colche], বলতে[bolte], লিখবে [likhbe], etc.

Combined inflections are the combination of two or more several Independent inflections, e.g., ছিলাম (করেছিলাম [korechilam]), ছিলেন (বলছিলেন [bolchilen], ছেন (খেলছেন [khelchen], লাম (করলাম [korlam], লেন (বললেন [bollen]), তেন (খেলতেন [khelten]), তাম (হাঁটতাম [hattam]), বেন (বলবেন [bolben]). In combined inflections, we observe that, most of the cases, an element from the Independent inflection Set-1 is appeared at last. For single length, the stem is often followed by additional য় [yo] or ও [o] which are not treated as inflection; rather, they are parts of stem word. e.g., খা[kha] >খাও, খায়[khay], খেয়েছিলাম [kheyechilam], নে[ne] >নাও [nao], নিয়েছিলাম [niyechilam]. We formulate our stemming algorithm as follows:

Step 1: For inflected words, the inflections (both independent and combined) are found examining the given word letter-by-letter starting from the end. Whenever an inflection is found in independent form, it is eliminated and a check is carried out whether it was a part of any combined inflection. If it is, the remaining part is also eliminated, e.g., in case of করলেন [korlen] and চলন[colon], our algorithm eliminates ন[na] first, then checks whether the eliminated ন [na] was a part of any combined inflection or not. For করলেন [korlen], it eliminates লে[le] (as it is a combined inflection) and leave কর [kor] as stem word, but in চলন [colon], the algorithm produces চল [col] as the stem word. This process continues for all other inflections. e.g.,

- করছিলাম [korchilam] >করছিলা [korchila] >করছি [korchi] >কর [kor] (Do)
- বললাম [bollam] >বললা [bolla] >বল [bol] (Talk)
- খেয়েছিলাম [kheyechilam] >খেয়ে [kheye] (Eat)
- নিয়েছি [niyechi] >নিয়ে [niye] (Take)

Step 2: If a word under processing reaches in its range of being a stem (Maximum length 3) and is ended with any diacritic marks (া, ি, ী, ু, ূ, ে, ৈ, ো, ৌ), the corresponding stem word can be retrieved easily by eliminating that diacritic mark only. e.g., করে [kore] will turn into কর [kor], বলি [boli] will turn into বল [bol]. For looking up dictionary entries [11], at the end of বল [bol], "া" is added. For single length stem or words ended with য়[yo] or ও[o], this approach will be ignored and at the end "ওয়া" [oa] is added with them, e.g., খাও[khao], খায়[khay] >খাওয়া [khawa] (Eat) etc.

Step 3: Now, we point on the first diacritic mark of the retrieved word from step 2, some transformations might require to get the correct stem word. e.g. transforming ে to া, ু to ো, ি to ে, etc. are required in খেয়েছিলাম [kheyechilam] >খেয়ে [kheye] >খেওয়া[kheoa] (খে) [khe] >খাওয়া[khoa] (Eat), নিয়েছিলাম [niyechilam] >নিয়ে[niye] >নিওয়া[nioa] (নি) [ni] >নেওয়া [neoa] (Take).

Step 4: Besides these rules, some special cases (i.e., exceptions) are also handled for incomplete stem words. e.g. আন [an] >আনা [ana] (Bring), আস [ash] >আসা[asha] (Come), এসেছিলাম [eshechilam] >আসা [asha], আয় [ay] >আসা[asha], এলেন [alen] >আসা [asha], গিয়েছিলাম[giyechilam] >যাওয়া [jaoa] (Go), খাচ্ছিলাম [khachhilam] >খাওয়া [khaoa] (Eat).

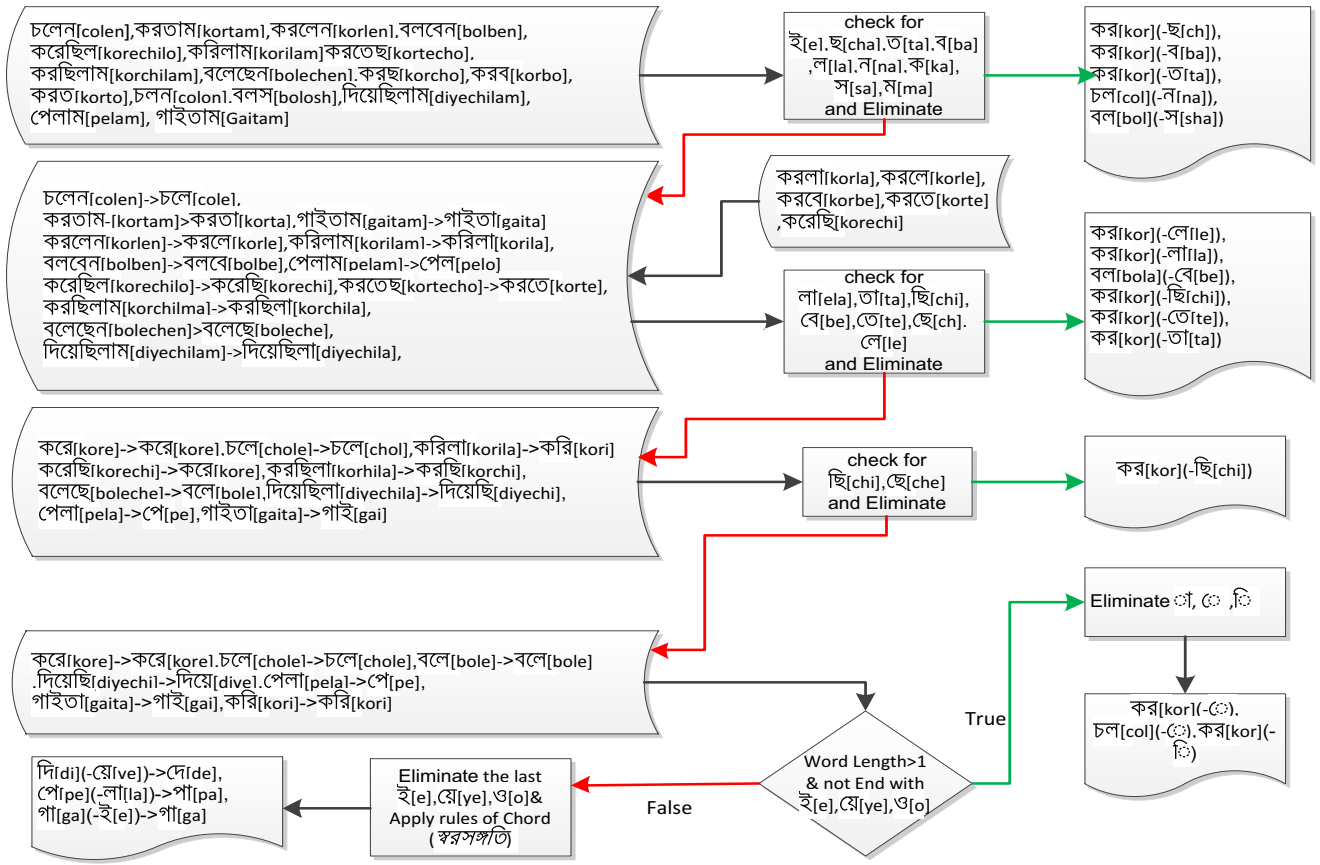

Fig. 1. Verb Inflection Elimination.

B. Algorithm for Noun Inflection

In noun inflections, we observe the following facts:

- Most of the noun inflections are used independently, e.g. গুলো [gulo]: গাছগুলো[gachgulo](Trees), দের[der]: ছাত্রদের [chatroder] (Students), তে[te]: নদীতে [nodite] (River), কে[ke]: মাকে [make] (Mother), etc. Some inflections following others may form combined inflections, e.g., গুলোতে [gulote]:বইগুলোতে [boigulote] (Books), দেরকে [derke]: অতিথিদেরকে[otithiderke] (Guests), etc. Single length stem and those words ended with vowel may have য়ের [yer] inflection (মায়ের [mayer], পায়ের [payer],ভাইয়ের [vaier], বউয়ের[bouer], বইয়ের [boier], etc.)

Based on the aforementioned observations, we formulate our stemming algorithm for noun as follows:

Step 1: As independent noun inflections are limited in number, they can be eliminated by scanning the given word letter-by-letter starting from the end. But, unlike combined verbal inflections, combined noun inflections do not follow any order of independent noun inflections. So, for a combined noun inflection, the order of eliminating its independent components is very important. The order of elimination, as shown in Fig. 2, is suitable for most of the combined noun inflections.

Step 2: Whenever the inflection is য়ের [yer], it checks whether the stem is single length or ended with vowel, based on that result, it is eliminated. e.g.মায়ের[mayer] >মা[ma] (Mother), but উভয়ের [uvoyer] >উভয় [uvoy] (Both).

Step 3: বৃষ্টির [bristir] >বৃষ্টি[bristi] (Rain), but মেষটির [meshtir] >মেষ[mesh] (Ram) - these type of variants are handled specially.

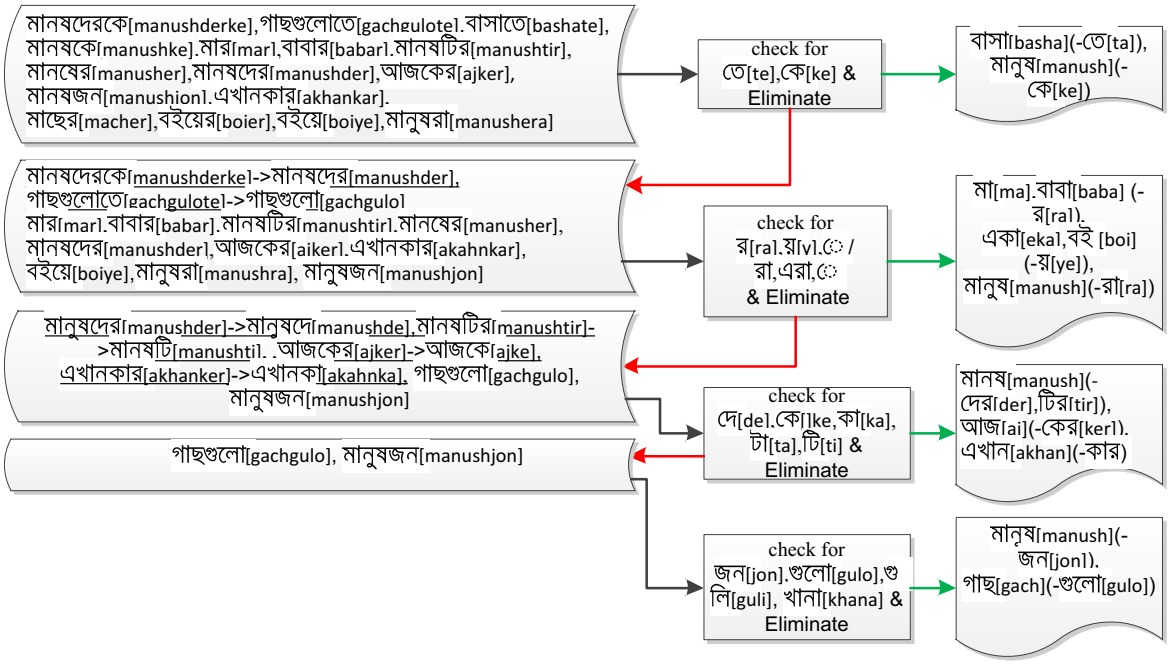

Fig. 2. Noun Inflection Elimination.

IV. Performance Evaluation

We have implemented our proposed stemming algorithms in PHP language. We have also used a PHP plug-in for Unicode to decimal conversion. We first conduct our experiments for the 20 categories of root words, listed in section III(A), and for all possible variations of verbs ending. The results observed are summarized below:

TABLE IV. EVALUATION RESULTS

Category	Input Size	Correct Stem	Accuracy
Verb	3000	2506	83%
Noun	1500	1325	88%

As the noun inflections are limited, they can be easily identified and eliminated with higher accuracy (88%) than verbal inflections (83%). These results are quite similar to the accuracy rate described in [6] 83% and in [8] ~89% using other techniques. Again, we took different datasets consisting of verb and noun inflections from two daily online Bengali newspapers – "The Daily Prothom Alo" and "bdnews24" and calculate the efficiency of our algorithm with respect to time. The following graphs describe the observations.

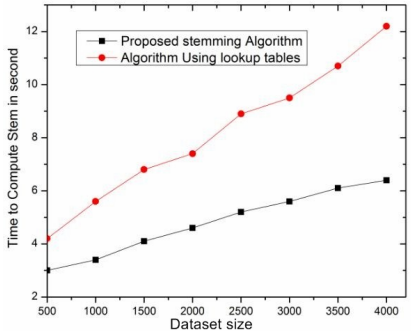

Fig. 3. Required time for Stemming Vs Dataset size.

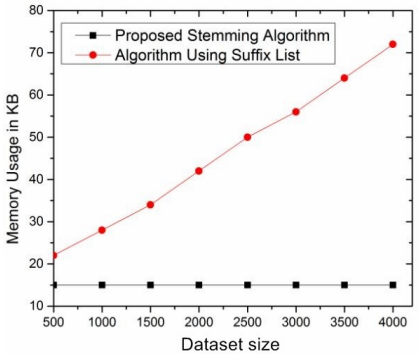

Fig. 4. Required memory for Stemming Vs Dataset size.

The computation time (Fig. 3) and memory (Fig. 4) required by our proposed rule based stemmer are much less compared to traditional algorithms using lookup tables. Furthermore, the time and memory usage gap between the two

increases with the input dataset size; our rule based algorithm takes almost half of the time and one fourth of the memory required by look up based algorithms, for 4000 datasize, as shown in Fig. 3 and 4, respectively. The main reason behind getting this nice result is due to the reduction of table lookup time in our algorithm and performing several algorithmic steps to strip of the suffixes within the programme, not by using external suffix list which is much memory consuming.

However, our proposed rule based stemmers fail to handle some irregular cases. Table V and VI list such irregular verbs and nouns, repectively.

TABLE V. SOME UNDETERMINED VERB STEMS

Input Word	Output	Correct Root
কচলালেন [kochlalen]	কচা [kocha]	কচলা [kochla]
উলটাও [ultao]	উল [ul]	উলটা [ulta]
কেনো [keno]	কেওয়া [keoa]	কেনা [kena]
আঁক [anko]	আঁা [an]	আঁকা [anka]
যাবেন [jaben]	যাবা [jaba]	যাওয়া [jaoa]
পৌঁছেছিস [pouchesish]	পৌঁা [pou]	পৌঁছা [pocha]
দিলেন [dilen]	দেলা [dela]	দেওয়া [deoa]
দিলাম [dilam]	দেলা [dela]	দেওয়া [deoa]
দিবেন [diben]	দেলা [dela]	দেওয়া [deoa]

TABLE VI. SOME UNDETERMINED NOUN STEMS

Input Word	Output	Correct Root
মেঝের [mejher]	মেঝ [mejh]	মেঝে [mejhe]
স্ত্রীয়ের [strier]	স্ত্রীয় [striyo]	স্ত্রী [stri]
ভ্রূয়ের [vruyer]	ভ্রূয় [vruyo]	ভ্রূ [vru]
ভোঁয়ের [vnoyer]	ভোঁয় [vnyo]	ভোঁ [vno]
পাত্রের [patrer]	পাত্রে [patre]	পাত্র [patro]
পত্রের [porter]	পত্রে [potre]	পত্র [potro]
হাতে [hate]	হা [ha]	হাত [hat]

VI. CONCLUSION

This paper develops a rule based stemming process that can extract stems from almost all possible verb and noun inflections from a given word list. The rules in the proposed algorithms are based on some observations of Bengali inflections. Even some of the incomplete verb stems can be detected by our algorithm .

Our algorithms are limited to work for words containing conjugated letters, only verb and noun inflections; the inflections for other parts of speeches are not considered. In future, we plan to develop an efficient algorithm for extarcting different parts of speeches separately from a given word list and apply our rule based stemmers on them.

ACKNOWLEDGMENT

This work was supported by Code Crafters International Bangladesh. Dr. Md. Abdur Razzaque is the corresponding author of this paper.

REFERENCES

[1] D. Bijal, and S. Sanket, "Overview of Stemming Algorithms for Indian and Non-Indian Languages" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 1144-1146.

[2] N. Saharia, K. M Konwar, U. Sharma, and J. K Kalita, An improved stemming approach using HMM for a highly inflectional language. Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science Volume 7816, 2013, pp 164-173

[3] S. Das and P. Mitra, "*A Rule-based Approach of Stemming for Inflectional and derivational Words in Bengali."* In Proceeding of the 2011 IEEE Students' Technology Symposium, IIT kharagpur.

[4] Md. Zahurul Islam, Md. NizamUddin and M. Khan, 2004-2007. "A Light Weight Stemmer for Bengali and Its Use in Spelling Checker," working papers 2004-2007, Center for Research on Bangla Language Processing, BRAC University, Dhaka, Bangladesh. Unpublished.

[5] N. UzZaman and M. Khan, *"A Comprehensive Bengali Spelling Checker."* In the Proceeding of the International Conference on Computer Processing on Bengali (ICCPB), Dhaka, Bangladesh. 2006.

[6] S. Dasgupta, V. Ng."Unsupervised morphological parsing of Bengali*."* Springer Science Business Media B.V. 2007, Lang Resources & Evaluation DOI 10.1007/s10579-007-9031-y.

[7] M. S. Islam, "Research on Bangla Language Processing in Bangladesh: Progress and Challenges." 8th International Language & Development Conference 23-25 June 2009, Dhaka, Bangladesh.

[8] S.Sarkar and S. Bandyopadhyay, "Design of a Rule-based Stemmer for Natural Language Text in Bengali." In Proceedings of the IJCNLP-08 workshop on NLP for Less Privileged Languages Hyderabad,India, pp.65-72.

[9] Willett, P, "The Porter stemming algorithm: then and now." Program: electronic library and information systems, 40 (3). pp. 219-223, 2006.

[10] P. Majumder, M. Mitra, S. Parui, G. Kole, P. Mitra, and K. Datta, "YASS: Yet Another Suffix Stripper." ACM Transactions on Information Systems, 2006.

[11] Dr. Hayat Mamud,*Uchhataro Shanirvor Bishudhho Vasha-Shikhha*. The Atlas Publishing House, Dhaka, Bangladesh, 2006.

[12] H. Ruth Thompson, "Bengali - A comprehensive grammar."Published by routledge, 2 park square, Miltonpark, United Kingdom, 2010.

[13] S. Khatun, *Bangla Academy Byabaharik Bangla Abidhan [Bangla Academy Functional Bengali Dictionary].* Bangla Academy, Dhaka, Bangladesh, 2012.