Third Lab Report

On

DATA MINING

Project Title: A Bangla Stemmer Based on Rules and Dictionary Matching

Course Number: CSE 4223
Course Title: Data Mining

Submitted To,

Dr. Kazi Masudul Alam,
Associate Professor,
Computer Science and Engineering Discipline,
Science Engineering and Technology School,
Khulna University, Khulna.

Submitted by,

Sirdarta Prashad Banik;	Student ID:170206
Ragib Mehedi;	Student ID:170208
Talha Al Junaid;	Student ID:170221
Chandan Sarder;	Student ID:170222

Year: 4th; Term: Second
CSE Discipline,
Khulna University, Khulna.

Computer Science and Engineering Discipline

Khulna University, Khulna

Date of Submission: 16-02-2021

Chapter I

Introduction

1.1 Introduction and Background

The Bengali language has more than 150k words in its vocabulary. These words grouped to form a sentence by following grammatical rules. Due to logical reasons, sentences use a different form of words derived from a root word, such as বলেছি(have told), বলি(tell), বলছি(telling), বলব(will tell), বলেছিল(told), বলত(told) have been derived from the root word বলা(tell). While working in Natural Language Processing (NLP) models and problems, these words not help much. The main focus of NLP problems is to achieve the result from fewer words. Solving this problem saves a lot of processing time and disk space. To solve this problem Stemming is used which is applicable to almost all languages. Stemming is a technique used to extract the root form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems. For example, the stem of the words ***eating, eats, eaten*** is ***eat***. The basic concept to the solution of the problem is to eliminate inflections from a given word to derive its stem word. There are many different types of stemming approaches like affix stripping, co-occurrence computation, dictionary look-up, longest suffix matching, probabilistic including natural language processing approaches [1]. Most of them 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 [1]. 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. For example, root word মানুষ(Man) has inflected or derivational forms মানুষগুলো(People), মানুষেরা(Men), মানুষদেরকে(People), মানুষজন(People) etc. This paper introduces a computationally inexpensive stemming algorithm, which combines suffix removal and dictionary look-up to produce finest output than all others stemming algorithm for Bengali language.

1.2 Motivation

Bengali is highly morphologically inflected language. So, the stemming task is more complicated than English. For English there are many renowned stemmers such as porter stemmer and Snowball stemmer which have developed based on various stemming algorithms [3]. But for Bengali there is scarcity of good stemmers. There is a Bengali stemmer in python, "Fatick Stemmer" and "rafi-kamal/Bangla-Stemmer". These stemmers are only rule based and hence can not produce accurate results due to Bengali complicated morphological structure. We have introduced a technique which use dictionary look-up along with rule-based approaches. It will increase the performance in many times.

1.3 Objective

To the best of our knowledge, we have designed a comparative good stemmer for Bengali language. The building process of the stemmer includes these steps-

- At first, we collect huge Bengali data from different newspapers.
- Then we process these data and make a dictionary of about 450k words including a large amount of nouns.
- We study Bengali grammar and investigates in Bengali morphology for a large set of inflections.
- Then we compute the stems algorithmically cutting down the inflections step by step from the end of the word along with dictionary matching.
- Compare our outputs with other stemmers outputs.

Chapter II

Literature Review

There are very limited researches that have been done in the field of Bangla language stemming. We have hardly found two papers and two stemming applications for Bangla language. They are discussed below-

2.1 Designing a Bangla Stemmer using rule based approach

MD Shahidul et.al. [2] have applied two rules to implement the stemming algorithm. They remove the long and short length bevokti(বিভক্তি) suffixes and bochon(বচন) suffixes from the end of the word.
They have used **rule1** to identify and eliminate bochon(বচন) suffixes as-

বৃন্দ/ মন্ডলী/ কুন্জ/ পুঞ্জ/ গুচ্ছ / সমুদয় / সমূহ/ বর্গ / রাশি / আবলি / থানা/ গুলি / রাজি / ইকর / খানি/ নিচয়/ গুলো / মালা/ যূথ/ সকল/ বলি/ দের / এরা / দিগ / পাল / দাম / কূল / টা/ রা/ সব / গণ/ টি/ আন -> ε

And **rule2** to identify and eliminate bevokti(বিভক্তি) suffixes as-

েছ/ িইয়েছ/ ায়/ াচ্ছি/ ায়েছি/ াচ্ছে / াতিস / ালাম / ালেন / াইলে/ াইবি / াবেন / াইতে / াতেন / াচ্ছ / াইলি / াতাম / াইবে / াইব/ ালি / ালে / াউক / াবো / াইস / ায়ো / াবে / াইও / াইয়/ াবি/ াতে/ াছে/ াক/ াও / াইয়েছিলাম/ াইতেছিলেন / াইয়েছিলাম/ াইতেচ্ছিস/ াইতেছিস/ াইয়েছিলেন/ াইয়াছিলেন/ াইতেছিলাম/ াইয়াছিলে / াইতিছিলে / াইতিছিলি/ াইয়েছিলি / াইয়াছিলি/ াইয়েছিলে/ াচ্ছিলাম/ াইতেছিলে/ াচ্ছিলেন / াইয়েছিস / াইয়েছিল/ াচ্ছিলে / াইয়াছিস / াইতেছেন / াইয়াছেন/ াচ্ছিলি/ াচ্ছিস / াইয়াছি/ াচ্ছেন/ াইতেছি/ াইয়াছি / াচ্ছেন / াইতেছি/ াইয়েছি/ াইয়াছে / াইয়েছে/ াইতেছে/ াহিতেছ / াইতেন/ াইলেন/ াইতাম/ াইতিস/ াইবেন/ াইলাম/ াইয়েছ/ াইয়াছ/ াইতাম / াইতিস/ াইবেন/ াইলাম / াইয়েছ/ াইয়াছ / িয়েছেন/ িয়েছ/ ি য়েছিস/ িয়েছি/ াইলি/ োচ্ছিল/ িতেছিল/ িয়াছিলে/ েছিলেন/ েছিলে/ িয়াছিলি/ েছিলি/ িয়াছিলা/ িবে/ িবি/ িয়া/ িবেন/ বেন / িয়া/ িস/ িয়েছেন/ িয়াছি/ াইল -> ε

They also make a do not stem table which contains such words that if these are stripped then these will lose its real meaning. Their work will not cover all possible rules. After all they have achieved about than 91% accuracy.

2.2 A Rule Based Bengali Stemmer

Md. Redowan et.al. [1] proposed a work that considers only two parts of speech noun and verb for stemming job. Bengali words are mostly inflected due to verbal and nominal inflections and they have a wide list of inflectional suffixes. 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. These inflections are given clearly in below table-

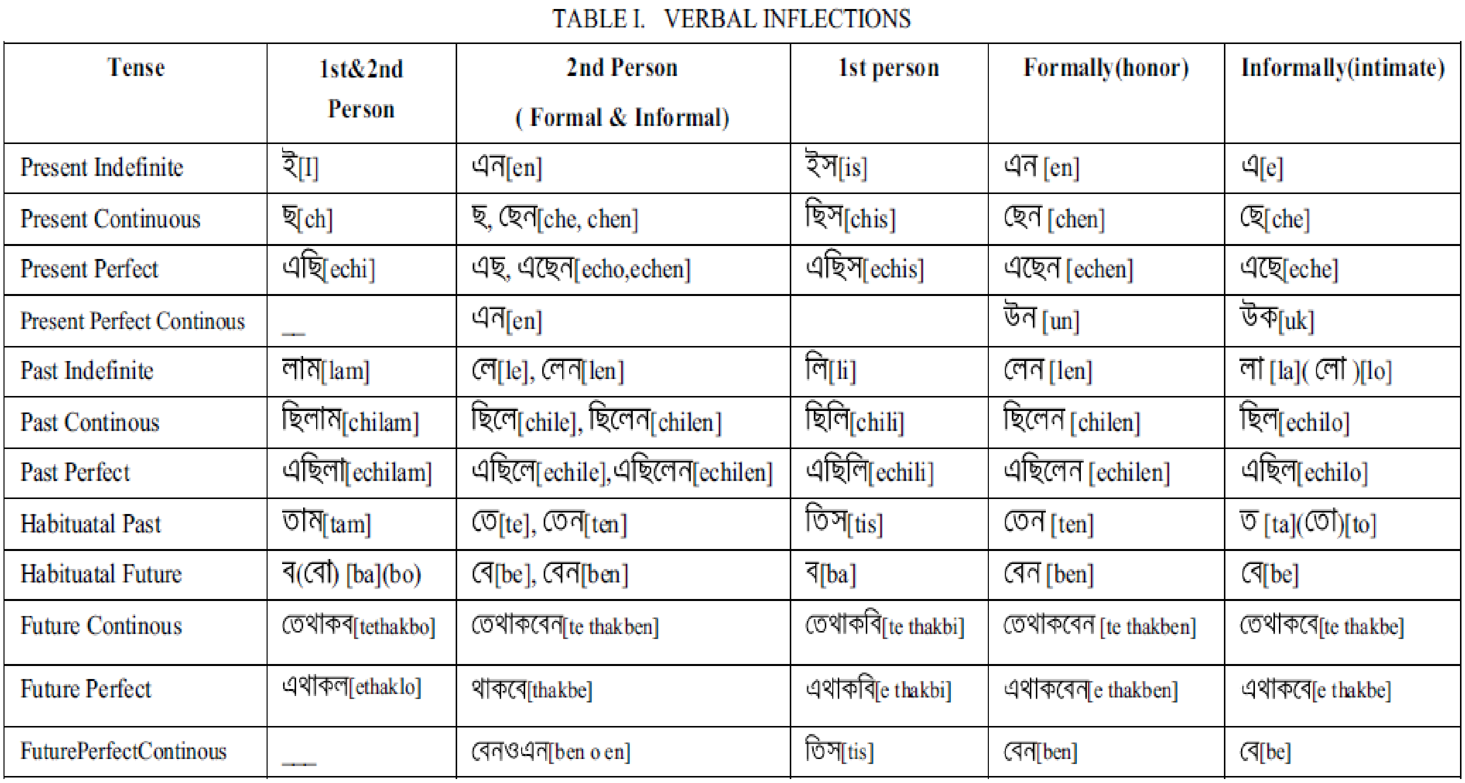

TABLE I. VERBAL INFLECTIONS

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 thak ben]	তেথাকবি[te thakbi]	তেথাকবেন [te thakben]	তেথাকবে[te thakbe]
Future Perfect	এথাকল[ethak lo]	থাকবে[thakbe]	এথাকবি[e thakbi]	এথাকবেন[e thakben]	এথাকবে[e thakbe]
FuturePerfectContinous	___	বেনওএন[ben o en]	তিস[tis]	বেন[ben]	বে[be]

Figure-1: Verbal inflection according to Md. Redowan et.al. [1]

On the other hand, noun inflections occur due to different cases like nominative, objective, genitive and locative. These cases also differ for singular and plural. These inflections are given clearly in below table-

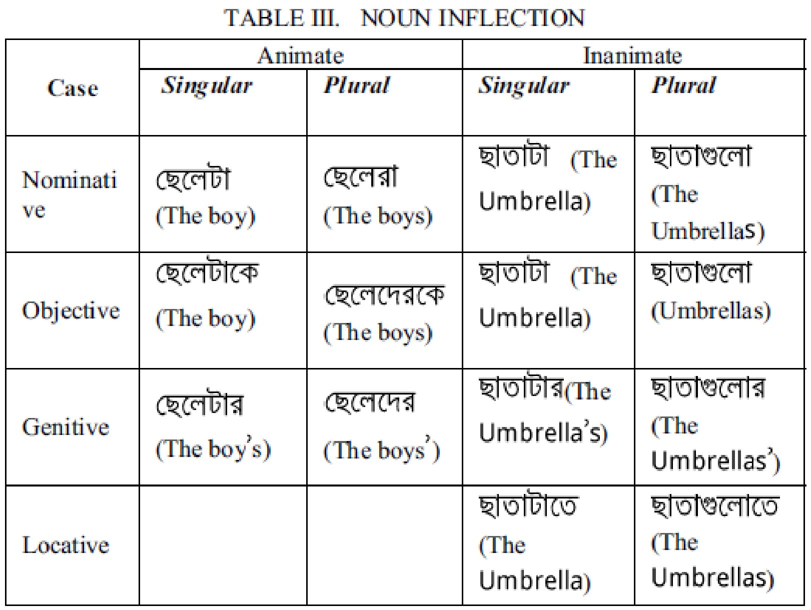

TABLE III. NOUN INFLECTION

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)

Figure-2: Noun inflection according to Md. Redowan et.al. [1]

This paper inflects some adjectives also. As the noun inflections are limited, they can be easily identified and eliminated with higher accuracy (88%) than verbal inflections (83%). In this paper the inflections for other parts of speeches are not considered. This paper’s proposed model is used in github project *BanglaKit Bengali Stemmer* [4].

2.3 bangla-stemmer 1.0 (A Python package to get stem of any inflected Bangla words)

This is a only Bangla stemmer available in official python website [5]. They have a github repository named as *Fatick Stemmer* [6]. This project uses the grammar rules of the stemmer of *Rafi Kamal* [7]. This project also referred to the another github project named as *BanglaKit Bengali Stemmer* [4]. BanglaKit Bengali Stemmer uses the algorithms of grammar rules from *Rafi Kamal's Stemmer* and *Mahmud's Stemmer* (algorithm of literature 2.2 [1]). All these stemmers are rule based. Most of the rules are about bivokti(বিভক্তি) and bochon(বচন). These rules are described in detail in our “**Bangla stemming Tool Deign**” section.

Fatick Stemmer: They have no paper about their project. From the github code it is seemed that their stemmer is also only rule based. They eliminate বিভক্তি and বচন which covers verb, noun, pronoun and some adjective words stemming [6].

Rafi Kamal: They have also no paper about their project. But from their project code we have noticed that they use total 55 suffixes of বিভক্তি and বচন to perform the stemming task [7].

Mahmud's Stemmer: Md. Redowan Mahmud et.al. [1] have a complete paper of their project named mahmud2014, partially used on github project *BanglaKit Bengali Stemmer*. Literature 2.2 have described about this project.

Chapter III

Proposed Method

3.1 Creating Dataset

3.1.1 Data Extraction from Different Bangla Newspapers

Newspaper is the largest source to collect a wide amount of vocabulary. In Bangla language there are five types of words. Newspapers contain these varieties types of words in a large amount than any other resources. There are many Bengali newspapers in Bangladesh. Among them, the Prothom Alo, the Kaler Kontho, the Daily Jugantor, the Daily Ittefaq, the Bangladesh Protidin, the Daily Jonokontho etc. are very common. In these newspapers, there are news of different categories such as, National, International, Sports, Culture, Education, Editorial, Economics, Finance and Banking etc. These news provide a huge number of vocabularies. We used crawler to extract data from newspapers. It is very tough to extract data from newspapers by reading their view page source. We collected 140000 data from the Daily Jonokontha of different categories. Unique words collected from the news were used as input in our project.

3.1.2 Data Categorization and Labeling

We extracted data from newspapers and categorized it according to types. The most common categories are National, International, Sports, Culture, Education, Editorial, Economics, Finance and Banking.

3.1.3 Data Preprocessing

Data preprocessing is a challenging task for our research. Because it will provide us a good collection of word vocabulary. Data cleaning is the vital part of preprocessing. We have found a quite large amount of garbage in our data. Among them, punctuations, symbols, special characters, mathematical operators, words from other languages, meaning less words etc. are very common. We have removed all of them from our data.

3.2 Data Analysis and Feature Extraction

3.2.1 Unigram

We collect all single unique words and identify the frequencies of these words. Single words are called as unigrams. After that, we sorted them according to alphabetic order and frequency wise. Then we use the alphabetically sorted unigrams to make a fulfill dictionary. In which all the words are situated in lexicographic order. Some of these words are used to test our stemming model.

3.2.2 Bigram

We collect all two words used in succession in a sentence and frequency of them. Units of two words used in succession in a sentence are called bigrams. We sorted them according to frequency and alphabet order. This is not used in our stemming project.

3.2.3 Trigram

We collect all three words used in succession in a sentence and frequency of them. Units of three words used in succession in a sentence are called trigrams. We sorted them according to frequency and alphabet order. This is also not used in our stemming project.

3.2.4 Sentences

We retrieved all sentences from the news. This is also not used in our stemming project.

3.3 Bangla Stemming Tool Design

We will discuss this section in tree main subsections. They are following-

- All Possible and Reliable Bangla Suffix Identification for Stemmer
- A Fulfill Bengali Dictionary Preparation
- Stemming with Suffix Patterns and Dictionary

3.3.1 All Possible and Reliable Bangla Suffix Identification for Stemmer

There are many suffixes that are used in Bengali language. We have studied the morphology of Bangla grammar and identified their structures to effectively stem them from root words [8]. We have used comparatively a large number of suffixes in our stemmer model than all other available models. The suffixes can be categorize in four different sections such as Number(‘বচন’), Bivokti (’বিভক্তি’), Verbal Bivokti (’ক্রিয়া বিভক্তি’) and Identifier (’পদাশ্রিত নির্দেশক’). They are described in detail below-

I. Rules for Number(‘বচন’):

Some suffixes are used to identify number. These are called Number(‘বচন’). These words are used to express plural form of singular form. These words are used as suffix and indicates plural form of the world. All these types of suffixes with their reduction rules are given in below list-

বচন - Rules	**Example**
গুলো → ∈	ইটগুলো → ইট

গন → ∈	শিক্ষকগণ → শিক্ষক
বৃন্দ → ∈	শিক্ষকবৃন্দ → শিক্ষক
মন্ডলী → ∈	সম্পাদকমন্ডলী →সম্পাদক
বর্গ → ∈	পন্ডিতববর্গ → পন্ডিত
কুল → ∈	কবিকুল → কবি
সকল → ∈	পর্বতসকল → পর্বত
সব → ∈	ভাইসব → ভাই
সমূহ → ∈	মনুষ্যসমূহ → মনুষ্য
আবলী → ∈	পুস্তকাবলী → পুস্তক
গুচ্ছ → ∈	কবিতাগুচ্ছ → কবিতা
দাম → ∈	কুসুমদাম → কুসুম
নিকর → ∈	কমলনিকর → কমল
পুঞ্জ → ∈	নক্ষত্রপুঞ্জ→ নক্ষত্র
মালা → ∈	মেঘমালা → মেঘ
রাজি → ∈	তারকারাজি → তারকা
রাশি → ∈	বালুরাশি → বালু
লোকেরা → ∈	লোক

II. Rules for Bivokti(‘বিভক্তি’):

Some suffixes are used that are identified as Bivokti (‘বিভক্তি’). These words are used after another words as suffix and co-ordinates with other words. The full list of Bivokti (‘বিভক্তি’) with their elimination rules are given in below list-

বিভক্তি - Rules	**Example**
তে → ∈	বলতে → বল
য় →∈	ঘোড়ায় → ঘোড়া
◌ের → ∈	মানুষের → মানুষ
র → ∈	মামার → মামা
রা → ∈	ছেলেরা → ছেলে
◌ের → ∈	ছাত্রের → ছাত্র
দের → ∈	ছাত্রদের → ছাত্র
কে→ ∈	হাসিমকে → হাসিম
◌ে → ∈	রসে → রস
◌েরা → ∈	ডাকাতেরা → ডাকাত

III. Rules for Verbal Bivokti('ক্রিয়া বিভক্তি'):

There are some suffixes similar to 'বিভক্তি' that only sit after verb roots known as Verbal Bivokti ('ক্রিয়া বিভক্তি'). They are varied according to tense and persons. We have worked on a wide list of Verbal Bivokti ('ক্রিয়া বিভক্তি') given in the table below-

ক্রিয়া-বিভক্তি - Rules	Example
াচ্ছেন → ায়	খাচ্ছেন → খায়
াচ্ছে → ায়	খাইয়াছেন → খায়
াচ্ছ → ায়	খাচ্ছ → খায়
াচ্ছিলাম → ই	খাচ্ছিলাম -> খায়
াচ্ছিলেন → ায়	খাচ্ছিলেন → খায়
াচ্ছিলে → ায়	খাচ্ছিলে → খায়
াচ্ছিল → ায়	যাচ্ছিল → যায়
াচ্ছিস → ায়	যাচ্ছিস → যায়
াচ্ছিলাম → ায়	যাচ্ছিলাম → যাই
াচ্ছি → ায়	যাচ্ছি → যায়
িচ্ছেন → েয়	নিচ্ছেন → নেয়
িচ্ছে → েয়	নিচ্ছে → নেয়
িচ্ছ → েয়	নিচ্ছ → নেয়
িচ্ছিলাম → েয়	দিচ্ছিলাম → দেয়
িচ্ছিলা → েয়	দিচ্ছিলা → দেয়
িচ্ছিলেন → েয়	দিচ্ছিলেন → দেয়
িচ্ছিলে → েয়	নিচ্ছিলে → নেয়
িচ্ছিস → েয়	দিচ্ছিস → দেয়
িচ্ছি → েয়	নিচ্ছি → নেয়
েয়েছে → েয়	নিয়েছে → নেয়
য়েছে → য়	হয়েছে → হয়
ইছে → য়	হইছে → হয়
েছে → য়	হয়েছে→হয়
েয়েছ→ আই	পেয়েছ→ পাই
িয়েছ→িয়	নিয়েছ→নিয়
য়েছ→য়	নিয়েছ→নিয়

IV. Rules for Identifier('পদাশ্রিত নির্দেশক'):

There is another type of suffixes in Bengali language commonly known Identifier ('পদাশ্রিত নির্দেশক'). We have reduced this suffix and all of them are given in the below list-

পদাশ্রিত নির্দেশক - Rules	**Example**
টি → ∈	বাড়িটি → বাড়ি
টা → ∈	বইটা → বই
খানা → ∈	কাপড়খানা → কাপড়
খানি → ∈	কলমখানি → কলম
গাছা → ∈	লাঠিগাছা → লাঠি
গাছি → ∈	চুড়িগাছি → চুড়ি

3.3.2 A Fulfill Bengali Dictionary Preparation

We needed to collect all possible words in the Bengali language. Because good dictionary is our main part of stemming tool along with grammatical rules. We look for Bangla words in Bangla Academy Dictionary but it has some limitations like it contains formal words, less nouns and less foreign words. So, we make our own dictionary from newspaper's unigram data though it has some limitations too but render good results to our research work. We clean the garbage data that may be caused by typing mistakes or font problems. We have tried to keep only the pure words as best as possible to make a fulfill dictionary.

3.3.3 Stemming with Suffix Patterns and Dictionary

We have designed a combined model of rules and dictionary matching. To do so first we analyze and identify all possible suffixes that sit after words. Then we search that whether these suffixes occur at the end of the words. If the suffixes are found then stemming process will start. At the same time, we search the word in the dictionary. If we found the word in the dictionary, we considered this word as the root word. Otherwise, there is no need to stem this word. Here are some examples of our stemming model.

Example of number(বচন) stemming:

Rules for number identification-

মন্ডলী/গুচ্ছ/দল/গুলো/গুলি/সমূহ/রাশি/মালা/পুঞ্জ/গুঞ্জ/সমুদয়/

সমূহ/বরগ/খানা/রাজি/সকল/যূথ/পাল/মালা/সব/কুল -> ε

Example: মেঘমালা -> মেঘ

Here, ‘মেঘমালা’ is used as plural form of ‘মেঘ’. ‘মালা’ is the suffix. It is used to make plural form of ‘মেঘ’. We also check that if ‘মেঘ’ is a valid word from the help of the dictionary. If we find that ‘মেঘ’ is a valid word then we finally stem it. Otherwise, we discard the stemming task and keep the word unchanged.

Similarly,

We have applied other rules for identifying

- **Bivokti(‘বিভক্তি)**
- **Verbal Bivokti(‘ক্রিয়া বিভক্তি’)**
- **Identifier(‘পদাশ্রিত নির্দেশক’)**

3.3.4. Pseudocode of Our Bangla Stemming Model Algorithm

Here is the pseudocode of our Bangla Stemming Model algorithm -

Step 1: Get all the dictionary words in a list.

Step 2: Get the words in a list that we want to stem.

Step 3: Store all four types of suffixes (‘বচন′, ′বিভক্তি′, ′ক্রিয়া বিভক্তি′, ′পদাশ্রিত নির্দেশক′) in different variables.

Step 4: Read each word from step 2 and check that any of the suffixes occurs at the end of this words. If matched then go to step 5 otherwise go to step 7.

Step 5: Fully remove the suffixes or replace the suffixes with necessary suffixes according to the rules of Bangla grammar. Then go to next step.

Step 6: After removing suffixes check the stemmed word is valid or not with the help of the dictionary. If the word is found in the dictionary then store the stemmed word and go to step 7. Otherwise, if the word is not found in the dictionary then do not stem the word and go to step 7.

Step 7: Store the stemmed word in a new list that may be modified or not according to the stemming rules.

Step 8: Print all the words from the new list.

Pseudocode with Example

Now we see the eight steps pseudocode with some examples,

Step 1: Suppose we get all the dictionary words in a list D.

Step 2: We take five words for stemming as, S=[‘আমার’, ’মামার’, ’সবুজ’, ‘কবিকুল’, ’খাচ্ছেন’]

Step 3: Store all four types of suffixes (‘বচন’, ’বিভক্তি’, ‘ক্রিয়া বিভক্তি’, ’পদাশ্রিত নির্দেশক’) in different variables.

Step 4: Check suffixes at the end of the five words,

Main Word	**Stemmed Word**	**Suffix**
আমার	আমা	র
মামার	মামা	র
সবুজ	সবুজ	ε
কবিকুল	কবি	কুল
খাচ্ছেন	খ	◌াচ্ছেন

We found suffixes for ‘আমার’, ‘মামার’, ‘কবিকুল’, ’খাচ্ছেন’. And no suffix for ’সবুজ’.

Step 5: Then we fully remove the suffixes or replace the suffixes with necessary suffixes according to the rules of Bangla grammar.

Stemmed Word with new suffixes	**Suffix of main word(Old suffix)**	**Rules** **Old suffix→New suffix**
আমা + ε = আমা	র	র → ∈
মামা + ε = মামা	র	র → ∈
সবুজ + ε = সবুজ	ε	No suffix, no rule
কবি + ε = কবি	কুল	কুল → ∈
খ + ◌ায় = খায়	◌াচ্ছেন	◌াচ্ছেন → ◌ায়

Step 6: Now we check whether the Stemmed are valid with the help of the dictionary.

আমা = Not Valid মামা =Valid সবুজ =Valid

কবি =Valid খায় =Valid

As আমা is not valid discard the stemming and store the main word আমার

Other words are stemmed finally as all of them are valid.

Step 7: Then we store all the stemmed words in a new list F.

F= [‘আমার’, ’মামা’, ’সবুজ’, ‘কবি’, ’ খায়’]

Step 8: Print all the words from the new list F.

আমার

মামা

সবুজ

কবি

খায়

3.4 Performance Measure

To measure the performance of our stemming model first we have compared our work with others. We have introduced new rules and algorithms in our research work. There is complete a python library named “Bangla stemmer 1.0”. It uses other’s stemmer’s rules and technique. We will compare our model with this model. To do this job we take a sample input of 30 words. Then we fed the data to both of that models and finally compare output with Bangla stemmer 1.0’s output. The table given below contains the comparison-

Comparison Table:

Main Word	Bangla Stemmer 1.0	Our Stemmer
কুমার	কুম	কুমার
শেয়ারবাজার	শেয়ারব	শেয়ারবাজার
থেকে	থেক	থেকে
টাকা	টাকা	টাকা
উত্তোলনে	উত্তোলনে	উত্তোলন
প্রিমিয়াম	প্রিমিয়াম	প্রিমিয়াম
খাচ্ছেন	খায়	খায়
খাইয়াছেন	খাইয়াছ	খাইয়াছ
খাচ্ছ	খায়	খায়
খাচ্ছিলাম	খায়	খায়
খাচ্ছিলেন	খাচ্ছি	খায়
খাচ্ছিলে	খাচ্ছিলে	খায়
নিচ্ছে	নেয়	নেয়
দিচ্ছিলাম	দিচ্ছি	দেয়
দিচ্ছিলা	দিচ্ছিলা	দেয়
দিচ্ছিস	দিচ্ছিস	দেয়

Main Word	Bangla Stemmer 1.0	Our Stemmer
নিয়েছে	নিয়	নিয়
হয়েছে	হয়	হয়
হচ্ছে	হচ	হয়
হইছে	হই	হয়
কোম্পানিগুলোর	কোম্পানি	কোম্পানি
ক্ষেত্রে	ক্ষেত্রে	ক্ষেত্র
বাধ্যতামূলক	বাধ্যতামূলক	বাধ্যতামূলক
করে	করে	কর
সালের	সাল	সাল
ডিসেম্বরে	ডিসেম্বরে	ডিসেম্বর
পাবলিক	পাবলিক	পাবলিক
করা	কর	কর
আমার	আমা	আমার
তোমার	তোমা	তোমার

Chapter IV

Conclusion

This paper develops dictionary look-up and rule based stemming model that can extract stems from almost all possible parts-of-speech inflections of Bangla language. The rules in the proposed algorithms are based on some observations of Bengali inflections. We have introduced four different rules for bivokti, verbal bivokti, number and identifier. These four types of rules use as suffixes after Bangla words and cover almost all the words of noun(বিশেষ্য), pronoun(সর্বনাম), verb(ক্রিয়া), adjective(বিশেষণ), conjunction(অব্যয়). A dictionary is developed by us which increase the performance of our model by identifying the root word’s validity. We have able to stem all types of words including a large number of nouns, foreign words and mixed words. After all we need to find out more complicated rules of Bangla grammar and improve the quality of the dictionary to achieve the accuracy near 100%.

References

[1] Mahmud, Md Redowan, et al. "A rule based Bengali stemmer." 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2014.

[2] Shakib, MD Shahidul Salim, Tanim Ahmed, and KM Azharul Hasan. "Designing a Bangla Stemmer using rule based approach." 2019 International Conference on Bangla Speech and Language Processing (ICBSLP). IEEE, 2019.

[3] https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8

[4] https://github.com/banglakit/bengali-stemmer

[5] https://pypi.org/project/bangla-stemmer/

[6] https://github.com/MIProtick/Bangla-stemmer

[7] https://github.com/rafi-kamal/Bangla-Stemmer

[8] https://drive.google.com/file/d/1Lz44Rw9btpgvBONTWYtDiagkkBwj_QNw/view