An Efficient Unicode based Sorting Algorithm for Bengali Words

Md. Ruhul Amin
Dept. of Computer Science and Engineering,
Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

Asif Mohammed Samir
Dept. of Computer Science and Engineering,
Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

Madhusodan Chakraborty
Dept. of Computer Science and Engineering,
Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

Md. Mahfuzur Rahaman
Dept. of Computer Science and Engineering,
Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

ABSTRACT
This discussion focuses on the sorting algorithm for Bengali language represented by Unicode character set. A few works has been done on this topics but no standard is set up yet to sort Bengali words. Some of these works are based on ASCII representation. To use ASCII, keyboard mapping is important because it is different for each country where the Unicode representation is fixed for all characters of various countries. So, Unicode representation is much more preferable than ASCII representation. In this paper we have discussed about an easy way to sort the Unicode Bengali texts. In our method, a mapping is used which simplified the sorting procedure. This method can sort any Unicode Bengali text and it is not keyboard dependent.

General Terms
Theoretical Informatics.

Keywords
Bengali Word Sorting, Unicode Bengali Sorting, Bengali Text Sorting

1. INTRODUCTION
Bengali is an eastern Indo-Aryan Language. It is the native language of Bangladesh, the Indian state of West Bengal and parts of the Indian states of Tripura and Assam. It is written with the Bengali script. [1] About 181 million people are the native speaker of this language and nearly 250 million people can speak Bengali in total. It is one of the most spoken languages (ranking sixth) all over the world. [2] It is the national and official language of Bangladesh and one of the 23 official languages recognized by the Republic of India. [1] It is the official language of the states of West Bengal and Tripura. [1] It is also a major language in the Indian union territory of Andaman and Nicobar Islands. [1] It was made an official language of Sierra Leone in order to honor the Bangladeshi peacekeeping force from the United Nations stationed there. [1] It is also the co-official language of Assam.

As the Bengali language is a rich and widely used language, it must have some standardization such as Bengali keyboard layout, Bengali character recognition, voice synthesis etc. But unfortunately we have advanced a very little in this regard. In a rapidly developing environment of computerization of Bengali language, one of the most important issues is Bengali text sorting. For the development of Bengali database systems, an efficient, versatile sorting algorithm is a must. There are some papers on this topic but none of them could set standard for sorting Bengali text. In this paper, we have shown the analysis of the previously proposed sorting algorithms and the comparison among the procedures to represent drawbacks, difficulties and limitations. Based on these observations we have proposed an algorithm based on Unicode to sort Bengali strings accurately, and the complexity is satisfying. The proposed algorithm is readable and very easy to code; hence it has the potential to be considered as standard algorithm for sorting Bengali strings. As Bangla Academy [3] is the national academy for promoting Bangla language in Bangladesh, we are following the Bangla Academy dictionary standard for our proposed method.

2. THE BENGALI LANGUAGE
Base Letters: In the written form of Bengali alphabets, there are 11 vowels and 39 consonants. When we use these alphabets, we call it base letters. The vowels are

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

The consonants are

ক খ গ ঘ ঙ চ ছ জ ঝ ঞ ট ঠ ড ঢ ণ ত থ দ ধ ন প ফ ব ভ ম য র ল শ
ষ স হ ড় ঢ় য় ◌ং ◌ঃ ◌ঁ

These are the base letters of Bengali language.

Modifiers: There two types of modifiers in Bengali, vowel modifiers and the consonant modifiers.

10 of the 11 vowels can be used as modifier to the consonants. We call them vowel modifier. They can never be used independently. Here is the list of vowel modifier:

Word	Vowel Modifier	Example
অ	–	কলম /kɔlom/
আ	া	কলাম /kɔlam/
ই	ি	পিঠা /piʈʰa/

ঈ	ী	জীবন /dʒībon/
ঊ	ু	তুলা /tulɑ/
ঊ	ূ	সূচী /suci/
ঋ	ৃ	বৃষ্টি /briʃṭi/
এ	ে	কেমন /kæmon/
ঐ	ৈ	হৈম /hojmo/
ও	ো	কোমল /komol/
ঔ	ৌ	শৌখিন /ʃowkʰin/

Table-01: Vowel Modifiers

Like the vowel modifiers, the consonants have some short forms when they used with other consonant. They are called –ফলা. Some of them are given below:

Word	Consonant Modifier	Example
ব	ব–ফলা	জ্বর /dʒɔr/
য	য–ফলা	জন্য /dʒonno/
র	র–ফলা	তীব্র /tibro/

Table-02: Consonant Modifiers

Compound Characters: When two or more consonant characters of Bengali alphabet used together, then they are called the compound characters. There are about 270 compound characters in Bengali. Some examples of compound characters are given below:

Word	Compound Character	Decompressed Form	No. of Alphabet Used
উজ্জ্বল /ujjbol/	জ্জ্ব	জ + জ + ব	3
বৃষ্টি /briʃṭi/	ষ্ট	ষ + ট	2
যুদ্ধ /dʒuddho/	দ্ধ	দ + ধ	2
ব্রাহ্মন /brɑmmon/	হ্ম	হ + ম	2
সম্বন্ধে /sɔmmɔndʰe/	ম্ব ন্ধ	ম + ব ন + ধ	2 2

Table-03: Compound Characters

3. DIFFICULTIES OF SORTING BENGALI TEXT

The problem associated with sorting of Bengali words are as follows-

- Bengali words should be sorted according to the Bangla Academy [4] standard. But unfortunately the Unicode for Bengali characters are not in Bangla Academy dictionary order. So, mapping is required to sort words uniquely.
- Compound characters (জ + ্ +জ + ্ +ব = জ্ব, জ + ্ + ঞ = জ্ঞ) make Bengali sorting complicated.
- In writing, vowel modifiers (ে +ক = েক, ক + া = কা) can precede or follow the base letter in Bengali words, but in computation it should be placed after the base letter for proper sorting.
- Unicode characters র, য়, ঢ়, ড় can be written in two ways. For example, ঢ় can be a single character ঢ় (\u09DD) or compound of ঢ + ় (\u09A2+\u09BC). These two cases should be considered as special case while sorting.

4. POSSIBLE SOLUTIONS

4.1 Method 1

M. Shahidur Rahman et al. [5] have proposed an alternative representation during computation. According to their proposal a dummy character is placed after the character, which does not have any modifier. Moreover, it is also considered that there would be no dummy character between the constituent parts of a compound character. Generally vowel modifiers can be typed before or after the characters but for this algorithm the modifiers are shifted after the character for the internal representation while computing. In case of compound characters, they are decomposed into their constituent components and stored accordingly. In Table-4 internal representation of few words are shown where □ represents the dummy character. To sort the words the relative order in the character set are arranged in the following way-

Null modifier < Vowel Modifiers < Vowels < Consonants

Input Word	Internal input Representation	Internal Representation of Sorted Output Word
কুসুম	ক ু স ু ম	ক □ ম □ ল া
নিলয়	ন ি ল □ য়	ক ু স ু ম
মৃণাল	ম ৃ ণ া ল	খ ো ক □ ন

খোকন	খ ো ক ◌ ন	ন ি ল ◌ য়
কমলা	ক ◌ ম ◌ ল া	ম ৃ ণ া ল
রেশমা	র ে শ ◌ ম া	র ে শ ◌ ম া

Table-04: Internal representation Of Words In [5]

This method has the following drawbacks:

- This work is based on ASCII based Bengali words.
- In this procedure, হসন্ত (্) is not considered.

4.2 Method 2

According to Mafizul Haque Khan et al.'s "An Efficient And Correct Bangla Sorting Algorithm" **[6]** a character is represented with two digit unique number for every letter of Bengali alphabet along with the vowel modifiers and the consonant modifiers. The letters and their corresponding numbers are given in Table-05. It is to be noticed that here 'আ ' is treated as a set of two characters that is 'অ + া '. The consonant modifiers are having the same number as their original consonants.

Character	Number
অ, ই-ঔ, ং, ঃ, ঁ	11-23
ক-ঙ, চ-ঞ	25-34
ট, ঠ, ড, ড়, ঢ, ঢ়, ণ , ◌	35-42
ত, থ, দ, ধ, ন, প, ফ, ব, ভ, ম	43-52
য, য়, র, র, ় ,ল, শ, ষ, স	53-61
া, ি, ী, ু, ূ, ৃ, ে, ৈ, ো, ৌ	71-80

Table-05: Representing A Bengali Character Of Two Digit

For vowel modifiers, wherever its position is (i.e. Left, Right or Down), its corresponding number will always be visual after the number of the letter over which it was applied. For Example, the words সা, সি, সু, সো, and সে change into numbers '6171','6174','6172','6179' and '6177'. For Compound characters 99 is added. For example, for the word- ক্ক = 'ক + ◌্ + ক', the number will be 259925. For সহ্য, it will be - 60619953. The difference of number of digit between the words গোধূলি (277946755773) and সংস্কৃতি (6021609925764372) is 4. So according to the algorithm, four zero's are appended at the end of the number representing গোধূলি. So finally the number becomes 2779467557730000.

Drawbacks-

- আ is considered only as a compound character of 'অ + া', which is incorrect.
- The sequence of characters does not maintain the sequence of Bangla Academy [4].
- Adding extra zero's at the end of the number with less digit which increases overhead.

4.3 Method 3

Shah Md. Emrul Islam et al. proposed a method to Sort Unicode Bengali Text Using Ancillary Maps. [7] In this method, the Unicode characters are mapped and given a Sort Weight. The structure of the Ancillary table is like the following:

Unicode	Sort Weight	Remarks
0985	01	
0986	02	
09BE	03	RM
0987	04	
09BF	05	LM
.	.	
.	.	
09B6	56	BL
09B7	57	BL
.	.	
09FA	89	
09BD	90	

Table-06: Structure of Ancillary Maps

For each word the mapped value are concatenated and a decimal point is added after two digits from the starting. Then it becomes a floating point number. By comparing all the floating point numbers, the list of words is sorted. For example, the word কানকো (Unicode Representation 099509BE09A8099509CB) gets the value 25.0346002519

The algorithm uses the decimal number system for determination of the value of a Bengali Word. There is an example for two Bengali words "কর্মচারী" and "কার্যকরভাবে", the decimal value becomes 25.005463510030035407 and 25.035463520025005400500349 15 respectively. Now it's easy to compare the two numbers. By this way a list of Bengali words can be sorted.

Drawbacks

- Adds extra complexity while converting a string to floating point number
- The range for the floating point decimal number may exceed if the length of the word is much longer which will arise round-off error

5. PROPOSED SOLUTION

5.1 Main Process

In our proposed method, at first we mapped all the characters that are used in Bengali text according to Bangla Academy [4] sequence.

At the first stage of word processing, an extra dummy character is being added after the base letter which has no modifier. But if there is a vowel modifier with the base letter then we place that modifier in place of the dummy character. For example, if we have a word "কলাম", then we make it ক+ ল+া+ ম. If there is no vowel modifier at the end of the last letter then the null modifier is not added. When we consider a compound character, then we get the base letter, a link character (্), and then the next character of the compound character and so on. For example, the word যুক্ত, we get য+ু+ক+্+ত. This procedure is also maintained for the letters with consonant modifier. Hence the alternative representation of the Bengali string is obtained.

In the next step, we generate a string of digits for each Bengali string using the map table (Table-07) to obtain a secondary representation. Then we sort these strings using ***MergeSort*** or any other efficient sorting algorithm using string comparison. Finally, we convert the secondary representations to its original Bengali strings.

The rules followed in our approach are-

1. Any character without any modifier is considered as character followed by null modifier.
2. Any character with by vowel modifier is considered as character followed by vowel modifier.
3. Any character with consonant modifier is considered as character followed by link character followed by consonant.
4. Any compound character is considered as character followed by link character followed by character.

In our algorithm the precedence of the Bengali character is maintained using the following rule:

Dummy/Null Character < Vowel Modifier < Consonant Modifier < Vowel < Consonant

5.2 Mapping

The whole mapping (Bengali character to a pair of digits) is given in the table below:

Unicode	Character	Value
09F9	৹	01
09BE	া	02
09BF	ি	03
09C0	ী	04
09C1	ু	05
09C2	ূ	06
09C3	ৃ	07
09C7	ে	08
09C8	ৈ	09
09CB	ো	10
09CC	ৌ	11
09CD	্	12
0985	অ	13
0986	আ	14
0987	ই	15
0988	ঈ	16
0989	উ	17
098A	ঊ	18
098B	ঋ	19
098F	এ	20
0990	ঐ	21
0993	ও	22
0994	ঔ	23
0982	ং	24
0983	ঃ	25
0981	ঁ	26
0995	ক	27
0996	খ	28
0997	গ	29
0998	ঘ	30
0999	ঙ	31
099A	চ	32
099B	ছ	32
099C	জ	33
099D	ঝ	34
099E	ঞ	36
099F	ট	37
09A0	ঠ	38
09A1	ড	39
09DC	ড়	40
09A2	ঢ	41
09DD	ঢ়	42
09A3	ণ	43
09CE		44
09A4	ত	45
09A5	থ	46
09A6	দ	47
09A7	ধ	48

Unicode	Character	Value
09A8	ন	49
09AA	প	50
09AB	ফ	51
09AC	ব	52
09AD	ভ	53
09AE	ম	54
09AF	য	55
09DF	য়	56
09B0	র	57
09B2	ল	58
09B6	শ	59
09B7	ষ	60
09B8	স	61
09B9	হ	62
09BC	়	63

Table-07: Our Proposed Map Values (According to the Bangla Academy Sequence [4])

5.3 Steps for Sorting Bengali Text

Step 1: At first, for each word a string is generated with dummy characters where needed. Here we represent the following words with their internal representations:

Input word	Internal Representation
আমার /amar/	আ ◌ম ◌া র
কলাম /kɔlam/	ক ◌ল ◌া ম
যুক্ত /ʤukto/	য ◌ু ক ◌্ ত
কলম /kɔlom/	ক ◌ল ◌ম
ধন্য /dhɔnyo/	ধ ◌ন ◌্ য
আম /aam/	আ ◌ম

Here the dummy character is '◌ '.

Step 2: Each of the Unicode character will be represented with string of two digits from the map. For the previous example, the representation of Unicode characters is given below:

Input word	Internal Representation	Representing with mapped value
আমার /amar/	আ ◌ম ◌া র	1401540257
কলাম /kɔlam/	ক ◌ল ◌া ম	2701580254
যুক্ত /ʤukto/	য ◌ু ক ◌্ ত	5505271245
কলম /kɔlom/	ক ◌ল ◌ম	2701580154
ধন্য /dhɔnyo/	ধ ◌ন ◌্ য	4801491255
আম /aam/	আ ◌ম	140154

Step 3: Now these generated strings can be sorted using any efficient sorting algorithm. After sorting the words of previous example we get the following order:

Input word	Internal Representation	Representing with mapped value
আম /aam/	আ ◌ম	140154
আমার /amar/	আ ◌ম ◌া র	1401540257
কলম /kɔlom/	ক ◌ল ◌ম	2701580154
কলাম /kɔlam/	ক ◌ল ◌া ম	2701580254
ধন্য /dhɔnyo/	ধ ◌ন ◌্ য	4801491255
যুক্ত /ʤukto/	য ◌ু ক ◌্ ত	5505271245

Step 4: In this step, we retrieve the sorted Bengali words by reverse mapping.

5.4 Algorithm

The proposed algorithm for sorting Bengali words is given below-

1. N ← Total no of words
2. **for** i ← 1 to N
3. For each word derive $MappedString_i$
4. Sort the Array containing MappedStrings
5. **for** i ← 1 to N
6. Reverse map the Bengali words from each $MappedString_i$

5.5 BengaliSort API

We have implemented our proposed technique in JAVA and created an API for sorting Unicode Bengali words using this algorithm. This API contains the classes BengaliSort, MapDefine, SortWords, MergeSort, QuickSort, RadixSort.

To sort words one of the methods below of class **BengaliSort** has to be called.

- void sort(String[] string)
- void sort(Collection<String> collection)
- void sort(Collection<String> collection, int fromIndex, int toIndex, boolean WithoutRepeatition)
- void sort(String string[], int fromIndex, int toIndex, boolean WithoutRepeatition)
- void sort(String string[], boolean WithoutRepeatition)
- void sort(Collection<String> collection, boolean WithoutRepeatition)

The link for the downloadable BengliSort API
http://www.msoden.com/

5.6 Complexity

Mapping and Reverse Mapping Function

N=total number of word in list or input list.

Mapping and reverse mapping is done in linear time. So, the Complexity of conversion from Unicode characters to integer string is O(N). The complexity of reverse mapping is also O(N).

Sorting

Complexity of MergeSort for N words = O(Nlog(N))

So, Total Complexity = O(N) + O(Nlog(N)) + O(N)
≈ **O(Nlog(N))**

6. COMPARISON

Runtime Comparison

Proced-ures ⇨ / No of Words ⇩	Proposed Approach	An Efficient And Correct Bangla Sorting Algorithm[6]	An Approach to sort Unicode Bengali Text Using Ancillary Maps[7]
10000	94 ms	93 ms	16 ms
50000	281 ms	296 ms	63 ms
100000	546 ms	609 ms	156 ms
500000	2730 ms	2980 ms	686 ms
1000000	5756 ms	5632 ms	1498 ms

Table-8: Runtime Comparison

7. OUTCOMES

Comparison in Table-08 shows that, the runtime for "An Efficient And Correct Bangla Sorting Algorithm [6]" is almost same as ours. But the proposed mapping does not maintain the sequence of Bangla Academy Dictionary [4]. For the approach described in "An Approach to sort Unicode Bengali Text Using Ancillary Maps[4]" has a much better runtime than proposed approach because it compares floating point of Bengali words. But this approach has a major problem of round-off error. Above all, the character sequence of the approach is not according to Bangla Academy dictionary [4]. The paper "*Bangla Sorting Algorithm: A Linguistic Approach*" [5] is ASCII based approach, so we did not consider this paper to compare with other approaches. We proposed a standard approach considering these drawbacks.

We have sorted the Bengali words using string comparison. This maintains the Bangla Academy [4] dictionary sequence. So far, string comparison has not yet been proposed by any other approach and our total complexity is satisfying. So, our algorithm can be considered as a standard one.

8. CONCLUSION

In this paper we have proposed an efficient and proper way to sort Bengali strings that conforms with the proper structure of Bengali words i.e. each modifier comes with a base letter. Our main effort was to maintain the right order according to the standard set by Bangla Academy while sorting and to preserve the general complexity of standard sorting algorithm. We have also tested the algorithm with more than 56,000 data taken randomly from Samsad Bengali-English Dictionary [15] in our algorithm and the output is completely in proper sequence as represented in Bangla Academy Dictionary. So this algorithm has the potential to be considered as the standard procedure for sorting Bengali strings based on Unicode character.

9. REFERENCES

[1] http://en.wikipedia.org/wiki/Bengali_language Retrieved 2011-05-11

[2] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers Retrieved 2011-05-11

[3] Bangla Academy: http://en.wikipedia.org/wiki/Bangla_Academy

[4] *Bangla Academy Bengali-English Dictionary*, First Edition June, 1994, Bangla Academy, Dhaka, Bangladesh.

[5] Rahman, Md. Shahidur and Iqbal, Md. Zafar, "*Bangla Sorting Algorithm: A Linguistic Approach*". Proceedings of International Conference on Computer and Information Technology, Dhaka, 18-20 December 1998, pp. 204-208.

[6] Mafizul Haque Khan, S M Rafizul Haque, Md. Sharif Uddin, Rahat Khan, A B M Tariqul Islam, "An Efficient And Correct Bangla Sorting Algorithm" 7th ICCIT, 2004 Page 125

[7] Shah Md. Emrul Islam and Muhammad Masroor Ali "An Approach to Sort Unicode Bengali Text Using Ancillary Maps", BUET, Dhaka.

[8] Cormen, Thomas and Leiserson, Charles and Rivest, Ronald: *"Introduction to Algorith*m", Prentice – Hall of India Private Limited, 1999.

[9] Ellis Horowitz and Sartaz Shani,: "Fundamentals of Computer Algorithm", Galgotia Publications Limited.

[10] Unicode Consortium- http://www.unicode.org/charts/PDF/U0980.pdf

[11] Mohammad, Kazi Din: "*Adhunik Bangla Byakoron O Rochona"*

[12] Rajesh Palit, Md. Abdus Sattar, "Representation of Bangla Characters in the Computer Systems", Bangladesh Journal of Computer and Information Technology, Vol. 7, No. 1, December, 1999.

[13] Masum, Md. Salahuddin, "*Study of Bangla Conjunctive Characters for Recognition*", B.Sc.Engg.Thesis, department of Computer Scince and Engineering, BUET, August 2001.

[14] Deitel and Santry "*Advanced Java 2 Platform*", Prentice Hall Publications.

[15] Knuth, Donald "*The Art of Computer Programming*", Addison-Wisely Publications, Boston

[16] Samsad Bengali-English Dictionary - http://dsal.uchicago.edu/dictionaries/biswas-bengali/

[17] Ishida, Richard - Bengali script notes http://rishida.net/scripts/bengali/