A Revised Unicode based Sorting Algorithm for Bengali Texts

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

ABSTRACT

This paper describes a sorting algorithm for Bengali texts which is one of the most vital tasks for Bengali Natural Language Processing. As Unicode is much more preferable than ASCII encoding, we need to use this representation for Bengali Language. But due to some distinct properties of Bengali Language, they cannot be sorted directly using the order in Unicode character scheme. A few works have been done on this topics – some of them are for ASCII encoding whether some are for Unicode. But still they have some drawbacks and still there is no standard to sort Bengali texts. In this paper, we have discussed about the previous approaches and proposing a revised and easier procedure to sort Unicode Bengali texts. We used a mapping to simplify the sorting process. The efficiency depends on the efficiency of the sorting algorithm. This method is able to sort any Unicode Bengali texts. It will also work for Unicode text of any language if we just change the mapping part. So the process is both keyboard and language independent.

General Terms

Theoretical Informatics

Keywords

Bengali Word Sorting, Bengali Text Sorting, Unicode Bengali Text Sorting, Bengali Linguistic Sort, Bengali Dictionary Sort, Bangla Academy Dictionary Based Sort.

1. INTRODUCTION

Bengali or Bangla is an Indo-Aryan language spoken predominantly in Bangladesh and in the Indian state of West Bengal and Tripura [1]. With about 250 million native and about 300 million total speakers worldwide, it is the second most spoken language in the Indian subcontinent, seventh most spoken language in the world by total number of native speakers and the tenth most spoken language by total number of speakers [1][2]. This language is derived from Sanskrit and hence appears to be similar to Hindi [3]. It is written left-to-right, top-to-bottom of page. Vocabulary of Bengali language is similar to Sanskrit and there are to some extent similarities with Latin. As it is one of the most spoken languages and it has some complexities in its structure, it becomes a fundamental necessity to have some standardization such as Bengali keyboard layout, Bengali character recognition, voice synthesis like speech to text or text to speech etc. Bengali text sorting is the first issue that need to be standardized first. There are some papers on this topic but still none of them could set standard for Bengali text sorting. In this writing, we have shown some analysis, drawbacks and limitations on the previous works. We also proposed a revised procedure that can be used as a standard procedure to sort Bengali texts. This procedure is easy to comprehend and implementation is so much easier in any programming language. It sorts the Bengali texts with Unicode representation according to Bangla Academy [4] standard. As Bangla Academy is Bangladesh's national language authority [5] and this is the national academy for promoting Bengali language in Bangladesh, we need to follow Bangla Academy to set standard for Bengali Linguistic works.

2. BENGALI LANGUAGE

Bengali language is written using the Bengali alphabet which is the 6th most widely used writing system in the world. The script shared by Assamese with minor variants and is the basis for the other writing systems like Meithei and Bishnupriva Manipuri [6]. The script has also been used to write Sanskrit in the region of Bengal.

2.1 Base Letters

There are 11 vowels and 39 consonants in the written form of Bengali alphabets. When we use these alphabets in full form, we call them base letters.

Independent Vowels (স্বরবর্ণ)

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

Consonants (ব্যঞ্জনবর্ণ)

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 ৎ ◌ং ◌ঃ ◌ঁ

2.2 Modifiers

There are two types of modifiers in Bengali alphabets – vowel modifiers and consonant modifiers.

Dependent Vowels or Vowel Modifiers (-কার)

10 of the 11 vowels are used as modifiers to consonants. They are called vowel modifiers and are generally known as -কার. They can never be used independently. Following is the list of vowel modifiers with examples:

Table 1. List of Vowel Modifiers with Examples

Vowel	Vowel Modifier	Example
আ	◌া	কা
ই	ি◌	কি
ঈ	◌ী	কী
উ	◌ু	কু
ঊ	◌ূ	কূ
ঋ	◌ৃ	কৃ

Vowel	Vowel Modifier	Example
এ	◌ে	কে
ঐ	◌ৈ	কৈ
ও	◌ো	কো
ঔ	◌ৌ	কৌ

Consonant Modifiers (-ফলা)

Like the vowel modifiers, some consonants have short forms when they are used with another consonant. They are called consonant modifiers and are generally known as -ফলা. Some of them are listed below with examples:

Table 2. List of Consonant Modifiers with Examples

Consonant	Consonant Modifier	Example
ন	ন-ফলা	যত্ন
ম	ম-ফলা	আত্মা
য	য-ফলা	জন্য
র	র-ফলা	প্রতি
ল	ল-ফলা	শুক্ল
ব	ব-ফলা	জ্বর

2.3 Compound Characters

When two or more consonant characters used together, then they are called compound characters. There are about 285 compound characters in Bengali [7]. Some examples of compound characters are listed below:

Table 3. Some Compound Characters with usage

Word	Compound Character	Decompressed Form	No. of Alphabets Used
উজ্জ্বল	জ্জ্ব	জ + ◌্ + জ + ◌্ + ব	3
উচ্ছ্বাস	চ্ছ্ব	চ + ◌্ + ছ + ◌্ + ব	3
দ্বন্দ্ব	দ্ব ন্দ্ব	দ + ◌্ + ব ন + ◌্ + দ + ◌্ + ব	2 3
বৃষ্টি	ষ্ট	ষ + ◌্ + ট	2
মুক্তি	ক্ত	ক + ◌্ + ত	2

2.4 Alphabetical order of Bangla Academy

Generally, we use the following alphabetical order everywhere:

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ

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 ৎ ◌ং ◌ঃ ◌ঁ

But ◌ং, ◌ঃ, ◌ঁ are used like a modifier and they cannot be used without any other alphabet. Though many compound characters are made up with consonant modifiers, they can also be written with conjunct character (◌্) between two consonants. To simplify these kind of complexities, Bangla Academy uses the following order for Bengali words in Dictionary:

অ আ ই ঈ উ ঊ ঋ এ ঐ ও ঔ ◌ং ◌ঃ ◌ঁ

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

We followed this alphabetic order to sort Bengali texts in our approach.

3. DIFFICULTIES TO SORT BENGALI TEXTS

The problems associated with sorting of Bengali texts are as follows:

- Bengali words should be sorted according to Bangla Academy [4] standard. But Unicode representation of Bengali alphabets are not in Bangla Academy Dictionary order. So, mapping is required to sort texts.
- Compound characters with consonant modifier or conjunct character make Bengali sorting more complex.
- Vowel modifiers can precede or follow the base letters in Bengali text, but the modifier should be considered after the base letter in computation for proper sorting.
- Unicode characters র, য়, ড়, ঢ় can be written in two ways – as a single character or as a compound character with ◌় character.
- Two vowel modifier ◌ো and ◌ৌ can be written as a single Unicode character or as preceding and following two modifiers.
- Ambiguity between র্য and র‌্য adds a bit more complexity in sorting Bengali texts. In both case, we get র + ◌্ + য but they are not same (র‌্য = র + ZWNJ + ◌্ + য).

4. PREVIOUS WORKS

Md. Ruhul Amin et al. [8] proposed an efficient Unicode based sorting algorithm for Bengali words. They have used null modifier which not mandatory. This approach cannot sort texts in the following situation:

Table 4. Situation cannot be solved by [8]

Word	Decompressed Form	Representation with mapped value
বসতি	ব ◦ স ◦ ত ি	520161014503
বস্‌তি	ব ◦ স ◌্ ত ি	520161124503
বস্তি	ব ◦ স ◌্ ত ি	520161124503

We actually get বস্তি = ব + ◦ + স + ্ + ZWNJ + ত + ি where ZWNJ is not mentioned in their process. So their algorithm will treat both বস্‌তি and বস্তি as same word.

Aamira Shabnam et al. [9] have described an easily comprehendible Unicode based sorting algorithm for Bangla words. They didn’t use any null modifier and used single digit mapping.

Table 5. Situation not handled by [9]

Word	Decompressed Form	Representation with mapped value
কলম	ক + ল + ম	255652
কলাম	ক + ল + া + ম	2556052

If the mapped string is sorted in lexicographical order, we will get কলাম before কলম which is not correct.

Aamira Shabnam et al. [10] have also described a faster approach to sort Unicode represented Bengali words. This paper also has the drawbacks of the previous one. In addition to this, the order mentioned in the discussion is different from Bangla Academy standard. They used just the regular sequence of Bengali alphabets.

Partha Sarathi Kar et al. [11] proposed an improved Unicode based sorting algorithm for Bengali words. It is a bit different from the previous approaches. They mapped each character and their modifier together and also mapped the joined letters. They used the mapping value according to the following order:

Base letter < Base letter with vowel modifier < Base letter with consonant modifier + Joint letter (according to order of each character)

There is about 285 joint letters [7] which we have mentioned earlier. The mapping for all alphabets and joint letters adds an extra overhead in this algorithm. Again, joint letters with more than two characters are not mapped here. So the words like উজ্জ্বল , উচ্ছ্বাস cannot be sorted using this algorithm.

5. PROPOSED METHOD

5.1 Assumptions

- Mapping is must as Unicode character set for Bengali is not sorted.
- We need to use same number of digits for mapping a letter or modifier to get rid of the drawbacks of [9] and [10].
- ZWJ (Zero-Width-Joiner) and ZWNJ (Zero-Width-Non-Joiner) should be considered while mapping and also while decompressing a word.
- It is important to maintain the alphabetic order or Bangla Academy to sort text according to Bangla Academy Dictionary.
- The precedence to follow Bangla Academy Dictionary order:

  ZWJ, ZWNJ < Vowel < Consonant < Vowel Modifier < Conjunct Character (্)
- We assume that, র, য়, ড়, ঢ় are made up with a single character, not a conjunct with ় character. ো and ৌ are also assumed as single modifier.

5.2 Mapping

Our proposed mapping scheme is listed below. We are proposing at least two digits for each letter or modifier.

Table 6. Mapping for our proposed method

Unicode Value	Character	Mapped Value
200C	ZWNJ	00
200D	ZWJ	01
0985	অ	02
0986	আ	03
0987	ই	04
0988	ঈ	05
0989	উ	06
098A	ঊ	07
098B	ঋ	08
098F	এ	09
0990	ঐ	10
0993	ও	11
0994	ঔ	12
0982	ং	13
0983	ঃ	14
0981	ঁ	15
0995	ক	16
0996	খ	17
0997	গ	18
0998	ঘ	19
0999	ঙ	20
099A	চ	21
099B	ছ	22
099C	জ	23
099D	ঝ	24
099E	ঞ	25
099F	ট	26
09A0	ঠ	27
09A1	ড	28
09DC	ড়	29
09A2	ঢ	30
09DD	ঢ়	31

Unicode Value	Character	Mapped Value
09A3	ণ	32
09A4	ত	33
09CE	ৎ	34
09A5	থ	35
09A6	দ	36
09A7	ধ	37
09A8	ন	38
09AA	প	39
09AB	ফ	40
09AC	ব	41
09AD	ভ	42
09AE	ম	43
09AF	য	44
09DF	য়	45
09B0	র	46
09B2	ল	47
09B6	শ	48
09B7	ষ	49
09B8	স	50
09B9	হ	51
09BE	া	52
09BF	ি	53
09C0	ী	54
09C1	ু	55
09C2	ূ	56
09C3	ৃ	57
09C7	ে	58
09C8	ৈ	59
09CB	ো	60
09CC	ৌ	61
09CD	্	62

5.3 Steps for Sorting

Step 1: Decompress each word into smaller parts like letter or modifier.

Table 7. First step for proposed method

Word	Decompressed Word
ক্যাঁট	ক ্ য া ঁ ট
ক্যাটালগ	ক ্ য া ট া ল গ
কাঁচ	ক া ঁ চ
কাচ	ক া চ
র‍্যাঁদা	র ZWJ ্ য া ঁ দ া
র‍্যাম	র ZWJ ্ য া ম
র‍্যাব	র ZWJ ্ য া ব
বসতি	ব স ত ি
বস্‌তি	ব স ্ ZWNJ ত ি
বস্তি	ব স ্ ত ি
বই	ব ই
বল	ব ল
বন	ব ন
উতরাই	উ ত র া ই
উৎরাই	উ ৎ র া ই
উত্তর	উ ত ্ ত র
কংস	ক ং স
কাংস	ক া ং স
কাঁক	ক া ঁ ক
কাক	ক া ক
আক্‌দ	আ ক ্ ZWNJ দ
আক্কেল	আ ক ্ ক ে ল

Step 2: Generate the mapped string with corresponding values for each letter and modifier.

Table 8. Second step for proposed method

Word	Decompressed Word	Mapped String
ক্যাঁট	ক ্ য া ঁ ট	166244521526
ক্যাটালগ	ক ্ য া ট া ল গ	1662445226524718
কাঁচ	ক া ঁ চ	16521521
কাচ	ক া চ	165221
র‍্যাঁদা	র ZWJ ্ য া ঁ দ া	4601624452153652
র‍্যাম	র ZWJ ্ য া ম	460162445243
র‍্যাব	র ZWJ ্ য া ব	460162445241
বসতি	ব স ত ি	41503353
বস্‌তি	ব স ্ ZWNJ ত ি	415062003353
বস্তি	ব স ্ ত ি	4150623353
বই	ব ই	4104
বল	ব ল	4147
বন	ব ন	4138
উতরাই	উ ত র া ই	0633465204

Word	Decompressed Word	Mapped String
উৎরাই	উ ৎ র া ই	0634465204
উত্তর	উ ত ্ ত র	0633623346
কংস	ক ং স	161350
কাংস	ক া ং স	16521350
কাঁক	ক া ঁ ক	16521516
কাক	ক া ক	165216
আক্‌দ	আ ক ্ ZWNJ দ	0316620036
আক্কেল	আ ক ্ ক ে ল	031662165847

Step 3: Store the mapped string either in the second column of a two-dimensional array or use a structure with two properties – text and mapped-string.

Step 4: Now sort the 2D array or the structure using any string sorting algorithm. It is important to sort lexicographically. Following is the sorted list of our examples:

Table 9. Sorted list with proposed algorithm

Word	Decompressed Word	Mapped String
আক্‌দ	আ ক ্ ZWNJ দ	0316620036
আক্কেল	আ ক ্ ক ে ল	031662165847
উতরাই	উ ত র া ই	0633465204
উত্তর	উ ত ্ ত র	0633623346
উৎরাই	উ ৎ র া ই	0634465204
কংস	ক ং স	161350
কাংস	ক া ং স	16521350
কাঁক	ক া ঁ ক	16521516
কাঁচ	ক া ঁ চ	16521521
কাক	ক া ক	165216
কাচ	ক া চ	165221
ক্যাঁট	ক ্ য া ঁ ট	166244521526
ক্যাটালগ	ক ্ য া ট া ল গ	1662445226524718
বই	ব ই	4104
বন	ব ন	4138
বল	ব ল	4147
বসতি	ব স ত ি	41503353
বস্‌তি	ব স ্ ZWNJ ত ি	415062003353
বস্তি	ব স ্ ত ি	4150623353
র‍্যাঁদা	র ZWJ ্ য া ঁ দ া	4601624452153652
র‍্যাব	র ZWJ ্ য া ব	460162445241
র‍্যাম	র ZWJ ্ য া ম	460162445243

Finally, we'll get the sorted texts in the first column of the 2D array or we can retrieve from the text properties of the structure. And this will be the same order which is maintained in Bangla Academy Dictionary.

5.4 Algorithm

The proposed algorithm for sorting Bengali texts is given below:

1. N ← Total Number of Words
2. foreach i in N
3. Derive Mapped Value for i-th word
4. Sort the words according to their mapped value in lexicographical order.

5.5 Complexity

For N words, the time complexity to generate the mapped value is O(N). If we use Merge Sort or any other efficient sorting algorithm for sorting, the complexity will me at most O(NlogN). So, the total time complexity for sorting N words will be

$$O(N) + O(NlogN) = O(NlogN)$$

5.6 Limitations

The only limitation in this procedure is the sorting order which is lexicographical. Without this this is to perfect algorithm for sorting Bengali text in a standard way.

6. TECHNICAL ANALYSIS

Using proposed algorithm, we can sort any text according to the order of Bangla Academy. None of the described procedures is able to sort the given examples in correct order. All the drawbacks discussed in the previous works section can be overcome by using our revised procedure. Following is a list of some examples that can depict the success of our proposed approach:

Table 10. Examples that can depict the success of our proposed solution

Input Words	কলাম	
	বস্‌তি	
	মাড়	
	উচ্ছ্বাস	
	বসতি	
	উজ্জ্বল	
	মাফ	
	কলম	
	বস্তি	
Sorting Status Using [8]		Cannot differentiate between বস্তি and বস্তি
Sorting Status Using [9]	…	কলাম will come before কলম which is wrong
	…	
	কলাম	
	কলম	
	…	
	…	
	…	
	…	
	…	
Sorting Status Using [10]	…	Mismatch with the sequence of Bangla Academy Dictionary [4]
	…	
	…	
	…	
	…	

	…	
	…	
	মাফ	
	মাড়	
Sorting Status Using [11]		Cannot sort উচ্ছ্বাস and উজ্জ্বল because they have compound characters with 3 alphabets
Sorting Status Using Proposed Approach	উচ্ছ্বাস	Successfully sorted according to Bangla Academy [4] Sequence
	উজ্জ্বল	
	কলম	
	কলাম	
	বসতি	
	বস্‌তি	
	বস্তি	
	মাড়	
	মাফ	

Our proposed algorithm is accurate and faster at a time. Again, no dummy characters are used in this approach and no reverse mapping is used.

7. CONCLUSION

Our main effort was to maintain the same ordering followed in Bangla Academy Dictionary. If the Unicode encoding scheme could be changed to the ordering of Bangla Academy, then we can avoid this problem easily. But using the current Unicode encoding scheme, we must use mapping to sort Bengali text. We've tested our algorithm for a large dataset with more than 62,000 data and it works fine within very short time. So this procedure can be set as the standard for sorting Bengali text according to Bangla Academy order.

8. REFERENCES

[1] https://en.wikibooks.org/wiki/Bengali

[2] https://en.wikipedia.org/wiki/Bengali_language

[3] Kenneth Katzner, 'The Languages of the World', Routledge, 1995.

[4] http://www.banglaacademy.org.bd/

[5] https://en.wikipedia.org/wiki/Bangla_Academy

[6] https://en.wikipedia.org/wiki/Bengali_alphabet

[7] http://forum.daffodilvarsity.edu.bd/index.php?topic=11714.0

[8] Md. Ruhul Amin, Asif Mohammed Samir, Madhusodan Chakraborty, Md. Mahfuzur Rahman, "An Efficient Unicode based Sorting Algorithm for Bengali Words"

[9] Aamira Shabnam, Debakar Shamanta Piklu, "An Easily Comprehendible Unicode Based Sorting Algorithm for Bangla Words"

[10] Aamira Shabnam, Tapashee Tabassum Urmi, Md. Saiful Islam, "A Faster Approach to Sort Unicode Represented Bengali Words"

[11] Partha Sarathi Kar, Shantanu Mandal, Labiba Jahan, "An Improved Unicode Based Sorting Algorithm for Bengali Words"

[12] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[13] *Bangla Academy Bengali-English Dictionar*y, First Edition June, 1994, Bangla Academy, Dhaka, Bangladesh.

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

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

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

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

[18] 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.

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

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

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

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

[23] Ishida, Richard - Bengali script noteshttp://rishida.net/scripts/bengali/