A Faster Approach to Sort Unicode Represented Bengali Words

Aamira Shabnam
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet 3114, Bangladesh

Tapashee Tabassum Urmi
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet 3114, Bangladesh

Md. Saiful Islam
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
Sylhet 3114, Bangladesh

ABSTRACT
Sorting Bengali words, a constituent part of Bengali language processing, Bengali data manipulation and Bengali database system comes up with a lot of challenges. A simple lexicographic ordering based on the Unicode representation does not yield the correct order of Bengali words as the character order in Unicode for Bengali differs from the order suggested by Bangla Academy. Besides, the presence of modifiers, compound characters, dual representation of some characters in Unicode as well as the precedence of vowels have made the situation even more complex. Our study aims to adapt the linguistic order for Unicode represented Bengali text while achieving maximum possible time and space efficiency. In this paper, we propose an approach to sort Bengali texts using popular algorithms with a slight modification in mapping so that it follows the linguistic order of the language and takes no extra memory. Also it shows a run time comparison with the previous works done on this topic.

General Terms
Natural Language Processing

Keywords
Bengali Word Sort; Unicode Bengali sort; Bengali Linguistic Sort; Bengali Dictionary Sort; Bangla Academy Sort

1. INTRODUCTION
[1] Bengali /bɛŋˈgɔːli/ or Bangla /bɑːŋlɑː/ (বাংলা), the language native to the region of Bengal, is a member of the Indo-Aryan group of the Indo-Iranian branch of the Indo-European language family. With about 220 million native and about 250 million total speakers, it is one of the most spoken languages, ranked seventh in the world. Hence, the computerization of Bengali language has become a fundamental necessity with the increased adaptation of natural language in the field of technology. For this reason, sorting of Bengali words, being one of the basic module for algorithmic analysis and database system, has developed into the pressing issue in its computerization for the past several years. A number of works has already been done on this issue. A fast and efficient approach using well-known sorting algorithms with slight modification has been proposed in this paper. This version of the algorithm cuts off a huge amount of processing time by restraining from the necessity to build strings with the mapped value while maintaining the accurate result even for the dual represented characters in Unicode. It also presents a comparative runtime study of all the works that have been done on this subject.

2. BENGALI LANGUAGE[2]
The Bengali language is written using the Bengali Script which comprises vowels, consonants, modifiers and compound characters.

2.1 Base Letters
There are about 11 vowels and 39 consonants in Bengali alphabet known together as Base Letter.

2.1.1 Vowels (স্বরবর্ণ)
অ আ ই ঈ উ ঊ
ঋ এ ঐ ও ঔ

2.1.2 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 of two kinds of modifiers in Bengali language.

2.2.1 Vowel Modifiers
The vowel modifiers are generally known as –কার. Out of 11 vowels, 10 are considered as modifiers.

া ি ী ু ূ ৃ ে ৈ ো ৌ

2.2.2 Consonant Modifiers
There are about six consonant modifiers in Bengali language. They are called –ফলা.

য-ফলা স্য র-ফলা ম্র ন-ফলা ন্ন
ল-ফলা ব্ল ম-ফলা ত্ম ব-ফলা ষ্ব

2.3 Compound Characters
Sometimes two or more consonant characters of Bengali alphabet are joined together and behave like a single character. They are known as the Compound characters. For example, জ্ঞ, ষ্ঠ etc. No one knows the correct number of compound characters in Bengali language. It is assumed that the number is 395.

3. PROPOSED SOLUTION

3.1 Behavior and assumptions

1. Two strings are compared according to how each string is read during computation.
2. Vowel modifiers follow a consonant.
3. A character followed by a consonant modifier is considered to be followed by ্ (হসন্ত) + the corresponding consonant of the modifier. i.e. ম্র will be read as ম+্+র.
4. A compound character is read with a ্ (হসন্ত) added between the base characters e.g. জ্জ্ব is read as জ+্+জ+্+ব where the last ্+ব holds for ব-ফলা at the end of জ্জ্ব.
5. The length of a string is determined by how the string is read i.e. যুক্ত (য+ু+ক+্+ত) has a length of 5.
6. The precedence of Bengali characters[3]: Vowels < Consonants < Vowel Modifiers < Consonant Modifiers.

3.2 Mapping

Table 1. Proposed Map Representation

Unicode Values	Characters	Mapped Values
0985	অ	0
0986	আ	1
0987	ই	2
0988	ঈ	3
0989	উ	4
098A	ঊ	5
098B	ঋ	6
098F	এ	7
0990	ঐ	8
0993	ও	9
0994	ঔ	10
0995	ক	11
0996	খ	12
0997	গ	13
0998	ঘ	14
0999	ঙ	15
099A	চ	16
099B	ছ	17
099C	জ	18
099D	ঝ	19
099E	ঞ	20
099F	ট	21
09A0	ঠ	22
09A1	ড	23
09A2	ঢ	24
09A3	ণ	25
09A4	ত	26
09A5	থ	27
09A6	দ	28
09A7	ধ	29
09A8	ন	30
09AA	প	31
09AB	ফ	32
09AC	ব	33
09AD	ভ	34
09AE	ম	35
09AF	য	36
09B0	র	37
09B2	ল	38
09B6	শ	39
09B7	ষ	40
09B8	স	41
09B9	হ	42
09DC	ড়	43
09DD	ঢ়	44
09DF	য়	45
09CE	ৎ	46
0982	ং	47
0983	ঃ	48

0981	ঁ	49
09BE	া	50
09BF	ি	51
09C0	ী	52
09C1	ু	53
09C2	ূ	54
09C3	ৃ	55
09C7	ে	56
09C8	ৈ	57
09CB	ো	58
09CC	ৌ	59
09CD	্	60
09BC	়	61
09D7	ৗ	62

3.3 String Comparison

String comparison is done by comparing two strings character by character using mapped values. For example, we take আমি (আ + ম + ি) and আমাকে (আ + ম + া +ক + ে) are two Bengali strings to be compared. In character by character comparison first mismatch is found in the 3rd position where ি occurs in আমি and া occurs in আমাকে. As the mapped value of া (50) is less than the mapped value of ি(51) then আমাকে will precede আমি. Similarly for খেলা (খ+ ে + ল + া) and খোলা(খ + ো + ল + া) খেলা will precede খোলা in their relative order.

There are also dual representations of some Bengali characters. For example: ‘ড়’ can be read both as a single character or assembly of two characters ‘ড’ and ‘়’. So if the machine reads ‘ড়’ as ‘ড’ followed by ‘় ’then in character by character matching we consider ‘ড’ and ‘় ’together as ‘ড়’, use the mapped value of ‘ড়’ and continue the naive string comparison approach. Other characters with dual representations are র(ব + ়), ঢ়(ঢ + ়), য়(য + ়), ো(ে + া), ৌ(ে + ৗ) which are also handled this way.

3.4 Algorithm

1. Map all the characters according to Bangla Academy
2. Sort the array of Bengali words with any comparison sort algorithm using the above string comparison method.

3.5 Complexity

Mapping is done in O (1).

Hence, the total complexity becomes the same as the sorting algorithm chosen.

3.6 Uniqueness

1. Fastest approach so far.
2. Reduces memory usage.
3. Processes the words only when comparing two strings resulting in a shorter processing time.
4. Avoids building new strings or using dummy character.

3.7 Limitation

This approach is limited to comparison sorts only.

4. Statistical Analysis

Tested on a computer with, Processor - Core i5

Clock-rate – 2.50 GHz RAM – 8GB

Used Language – JAVA

Used IDE – Netbeans 8.0.2

Used Algorithm – Merge Sort

Comparative Study between the Previous Algorithms and the Proposed Algorithm for Random Words

Table 2. Input Vs Runtime in millisecond

Algorithm	Input (Strings)			
	10000	50000	100000	500000
Proposed Algorithm	15	47	109	844
An easily comprehendible Unicode based sorting algorithms for Bangla words[4]	312	453	640	2890
An efficient Unicode based sorting algorithm[5]	94	281	546	2730
An Efficient And Correct Bangla Sorting Algorithm[6]	93	296	609	2980
An Approach to sort Unicode Bengali Text Using Ancillary Maps[7]	16	63	156	686

It is clearly visible that the proposed approach takes less time than any other algorithms.

5. CONCLUSION

The proposed approach assures to sort Bengali words faster than any other algorithms so far while maintaining the proper linguistic order set by Bangla Academy. This is why, It has the potential to be considered a standard sorting algorithm for Bengali Language.

6. REFERENCES

[1] Bengali language, Wkipedia: https://en.wikipedia.org/wiki/Bengali_language Retrieved Aug 01, 2015

[2] Bengali alphabet বাংলা Bangla (Bengali): http://www.omniglot.com/writing/bengali.htm Retrieved August 01,2015

[3] বাংলা একাডেমি: http://www.banglaacademy.org.bd/ Retrieved August 01,2015

[4] Shabnam, Aamira, and Debakar Shamanta Piklu. "An Easily Comprehendible Unicode based Sorting Algorithm for Bangla Words." *International Journal of Computer Applications* 79.5 (2013): 27-31.

[5] Amin, Md Ruhul, et al. "An Efficient Unicode based Sorting Algorithm for Bengali Words." *International Journal of Computer Applications* 24.7 (2011).

[6] Khan, Mafizul Haque, et al. "An Efficient and Correct Bangla Sorting Algorithm." *7th ICCIT* (2004): 125. K. Elissa, "Title of paper if known," unpublished.

[7] Islam, Shah Md Emrul, and Muhammad Masroor Ali. "An Approach to Sort Unicode Bengali Text Using Ancillary Maps." *Asian Journal of Information Technology* 4.10 (2005): 890-894.

[8] *The Unicode Standard 4. 0, copyright 1991-2003, Unicode, Inc.*

[9] *Bangla Academy Bengali-English Dictionary*. Bangla Academy, 1994.

[10] Mohammad, Kazi Din. "Adhunik Bangla Byakoron O Rochona." (1999).

[11] Thomas Cormen, Charles Leiserson, and Ronald Rivest: "Introduction to Algorithm", Prentice – Hall of India Private Limited, 1999.

[12] Dietel, Paul. *Java how to program*. PHI, 2009