Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Handle Bengali Numeral(s) in Rule based Bengali Spell Checker

Arghya Pal

Department of Computer Science & Technology,
Goa University, India

Abstract— In this paper I have made an attempt to plug in a sensor which handle the numeric character(s) for a rule based Bengali spell checker. The existing Bengali spell checker suggests discarding the numerals blindly and considering the rest part. But numerals play a vital role if they are seated at a proper position. So, before discarding them blindly why not we check their position, number of times they have occurred etc. The algorithm described below will handle Bengali numerals in a Context Sensitive manner. However this work is domain dependent in nature though we can modify it for any particular language.

Keywords— Numeric Character, Bengali, Bangla, Typographical error, Spell checker

I. INTRODUCTION

Numeric character (or say numerals) plays a very crucial role when we are trying to implement a rule based spell checker for any language. If the user enters a numeric value along with a text, then it is not a good idea to blindly discard the character(s) (i.e. numeric value(s)) all the time and generate suggestion(s) with the rest of the valid character sequence. Besides we can consider the position of the numeric character, font, and relative (or say neighbour) characters which leads to generate much more sophisticate suggestions for user. A typical spell checker in English should consider it a valid word if we enter a text like "1st" and assume that the user enters "1st" instead of first. Another example could be something like "21st February" which is the short form of "Twenty first February". Whereas "home1" or say "ho1me" are illegal words (if we only take a word and give it to spell checker to check the spelling).

In this paper I have made an attempt to create a rule based Bengali numeric character sensor for a Bengali spell checker in a domain dependent way, so if we enter any numeric character other than Bengali character it will be discarded or considered as an error. In Indian languages especially in Bengali, very less work has been done for numeric character recognition. On the other hand there is lot of both literature and application (online and offline Software) are available to implement a rule based spell checker in Bengali.

II. RELATED WORK WITH THIS ISSUE

Dr. Sivaji Bandyopadhyay[1] proposed an approach to get rid of, if the word starts with numerals. In this approach the numerals are just bypassed and the unit that are present after this numeral are considered. Unfortunately there is no prominent work till date to handle numeral in Bengali rule based spell checking.

III. DEALING WITH NUMERALS

Numerals play a very crucial role if they seat at a proper position. For example ১ম or ১নং is a valid (similarly in English 1st, 2nd) .But on the other hand কণ1জ should be a possible mistyping of কণৌজ "Kanauj" and considerd as a Typographical error described by Kukich. 1992a[2]. It mainly denotes the abbreviation of dates, street number in day to day scenario etc. After studying commercial advertisements, news paper, question papers the following situations have been obtained as give in TABLE 1:

TABLE I

OCCURRENCES OF NUMERALS AT DIFFERENT POSITION AND THEIR MEANING

Serial number	Input word	Use
1	১ম , ১নং , ২য়	To denote Brand quality, Topper in a Class, Street number, counts of day.
2	১লা, ২য়	To denote date, relative order of some sequence, Publication number of some magazine or journal.
3	১, 0	Pure numeral

4	১দিন	Day count
5	১২৩৪৫৬৭৮৯০	A phone number, Pin code, Memo number, Some Account number
6	0তম	Denote zero count

Numeral group

This array consists of all valid Bengali digits=[১,২,৩,৪,৫,৬,৭,৮,৯,০];

Procedure for dealing with numerals

Step 1: Take the input Text in Bengali;
If the Text starts with a " "(i.e a space,or,a tab) discard this " " and go ahead with rest Text

Step 2: Initialize a variable count=0and count1=0;
set a pointer NOW at the first character to the input text;

Step 3: While(NOW.value.isBelongstoNumeralGroup()==true){ //NOW.value will return the value at that position
// isBelongstoNumeralGroup() will return Boolean true if the character belongs to Numeral group array
count++;
NOW=next->NOW;//Increment the pointer position }

Step 4: If(count==0){call procedure Furthurecheck(); }

Step 5: Else If(count==1){call procedure ExactMatch(); }

Step 6: Else(count>1){
If((NOW.next.value.equlas()=="ত" && NOW.next.next.value.equals()=="ম") || (NOW.next.value.equlas()== "শ")) {Increase NOW by two position and
check If(NOW.value.equls()== " "){
Flag okey; // letters after space,tab will automatically be discarded by the spell checker }
}
Else If((NOW.next.value.equlas()=="ম"){take it as a typographical error and insert "ত" between numerals and "ম" then generate suggestion accordingly}
Else{If(NOW.next.value.equlas()==" "){1. Consider it as a phone number or a zip code etc
2. Flag Error to the user by printing "Did You Mean a dd/mm/yy ?"}} }

Step 7: If the character is not a Bengali numeral and nor a valid Bengali character it will be handled by the spell checker accordingly.

Procedure Furthurecheck

Step 1: go on increasing the pointer NOW If the value at that particular position of the pointer is not a numeral.

Step 2: Else{while(NOW.value.isBelongstoNumeralGroup()==true){ count1++;
NOW=next->NOW; //Increment the pointer position }}

Step 3:Suggest the spell checker to take it as a Typographical error and replace it with every possibilities.

Procedure ExactMatch

Step 1: If the input is ১or৫ or৭or৮or৯or০ followed by ,ম and " " ,or, নংand " ", Flag okey to the spell check

Step 2: Else If the input is ২ or ৩ followed by য়and " ", or, নংand " ", Flag okey to the spell checker

Step 3: Else If the input is ৪ followed by র্থand " ", or, নংand " ",Flag okey to the spell checker

Step 4: Else If the input is ৬ followed by,ম and " ",or, ঠ and " ", or, নংand " ", Flag okey to the spell checker

Step 5: Else Flag Error and replace the letter with neighbor key(say for example ১ with ঔ,ৌ,ৈ,ঐ etc.) letter to handle typographical error or, discard the letter and generate suggestion with rest of the inputs.

Possible Inputs and Outputs (or say suggestions) are listed in TABLE II

TABLE III
OCCURRENCES OF NUMERALS AT DIFFERENT POSITION AND THEIR MEANING

Serial number	Input word	Original Meaning in Bengali	Desired Output/Suggestion(s)	Possible type of error
1	১ম , ১নং	প্রথম "first"	প্রথম "Pratham"	Seems ok.
2	ম১	Error	মৌ "Mou", ১ম , প্রথম , ১নং	Transposition error for ১ম, Substitutionfor মৌ
3	১রাম	Error	১ম "1st", রাম "Rama", আরাম "AAram (Relax)"	Insertion error of ১

4	রা১ম	Error	১ম "1st", রাজা "Raja(king)", আম "AAm(Mango)", রাম "Rama"	Insertion error of ১
5	রাম১	Error	রাম "Rama"	Insertion error of ১
6	কণ1জ	Error	কণৌজ "Kanauj"	Insertion error of 1
7	১১ম	Error	১১তম , ১ম	Deletion error of ত
8	0	Valid and Error	0তম ,দ ,জ, ধ, ঝ	Typographical error
9	1রাবত	Error	ঔরাবত "Oyrabot", পারাবত "Parabot"	Typographical error

IV. IMPLEMENTATION AND RESULTS

I have implemented the above mentioned procedure with Java NetBeans 7.4 as front and phpMyAdmin (mysql) as back end Lexicon. The procedure has been tested on 1000 input from online Bengali news paper http://www.anandabazar.com/ , http://bartamanpatrika.com/, add agency like www.bengalimatrimony.com/ etc and the accuracy is 90%. Accuracy of Numeral is= Number of Correct output/ Total No. of Input text with numeral. The error 10% is due to the period (like ".", ";" etc in between the Text), small size of lexicon.

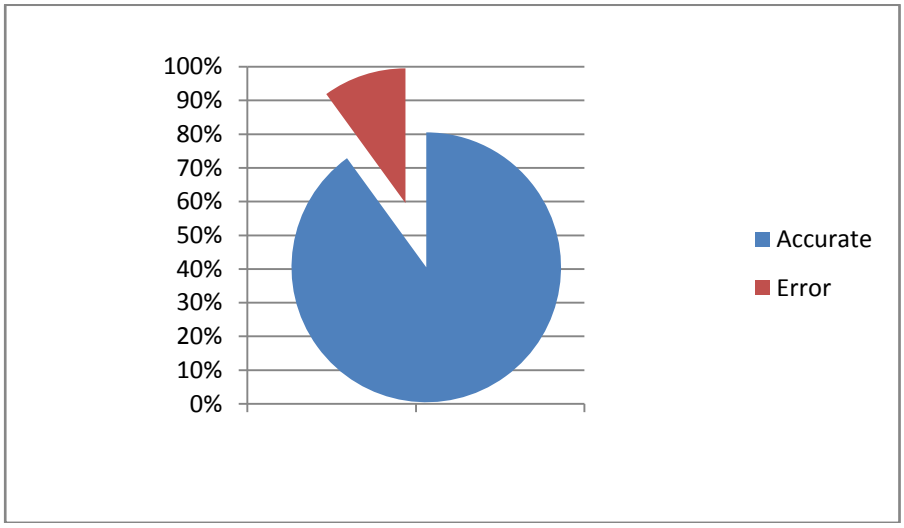

Fig 1: Accuracy and Error

IV. CONCLUSIONS

In this article I designed an light weight algorithm for checking the Bengali numerals. The algorithm will work as a plug in to the existing Bengali spell checker. Which on the other hand increase the accuracy of the existing spell checker. The accuracy of the algorithm can be improved by introducing more rules to the main procedure. The future work is to use this algorithm with more dataset from the news papers.

REFERENCES

[1] Dr. Sivaji Bandyopadhyay, *DETECTION AND CORRECTION OF PHONETIC ERRORS WITH A NEW ORTHOGRAPHIC DICTIONARY*, Computer Science and Engineering Department, Jadavpur University, ilidju@cal2.vsnl.net.in.

[2] Karen Kukich. 1992a. Spelling correction for the telecommunications network for the deaf. Communications of the ACM, 35(5):80-90, May

[3] Daniel Jurafsky & James Martin,*SPEECH and LANGUAGE PROCESSING*,Second Indian Reprint 2003

