IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-9 E-ISSN: 2347-2693

Contribution of 'Addak'and 'Bindi' in Non word Error Pattern analysis of Punjabi Typed Text

Meenu Bhagat

Dept. CSE, Punjab University SSG Regional Centre Hoshiarpur, Hoshiarpur, India

*Corresponding Author: meenubhagat@yahoo.com Tel.: +00-12345-54321

Available online at: www.ijcseonline.org

Received: 22/Aug/2017, Revised: 08/Sep/2017, Accepted: 19/Sep/2017, Published: 30/Sep/2017

Abstract— Error pattern analysis of a language is helpful in language related technology, such as creation of Natural Language Interfaces Spell Checker and Corrector, Machine Translation, Optical Character Recognition, Spell Checker and Corrector etc. It includes analysis of various types of errors (insertion, deletion, transposition, substitution, run-on, split word error) Positional analysis, Word length effects, Phonetic errors, First position error analysis, Keyboard effects etc. This paper mainly focuses on the contribution of 'Addak' and 'Bindi' in Statistical Error analysis of Punjabi typed text. It also compares results of insertion and deletion error results of 'Addak' and 'Bindi'and its overall contribution in total number of errors. It also presents a brief overview of difficulties in automatic text error correction in Punjabi Typed Text. This paper is based on the analysis done on 20000 misspelled words generated by typists.

Keywords: Addak, Gurumukhi, Non-word, Bindi.

I. INTRODUCTION

Chaudhuri and Kundu [1] have done a detailed analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based spellchecker for Bangla text. Pollock and Zamora [2] aimed at discovering probabilistic tendencies, such as which letters and position within a word are most frequently involved in errors, with the intent of devising a similarity key based techniques. Church and Gale[3] have done a Probability scoring for spelling correction. Kukich[4] has discussed the various techniques for automatically detection and correction of misspellings and the various factors affecting the spelling errors patterns of words in English. Damerau [5] worked on a technique for computer detection and correction of spelling errors in English language. Morris and Cherry [6] devised an alternative technique for using trigram frequency statistics to detect errors.Yannakoudakis and Fawthrop [7-8] sought a general characterization of misspelling behaviour. Wagner [9] was the first one to introduce the notion of applying dynamic programming techniques to the spelling correction problem to increase computational efficiency.

Section I contains a brief overview of work done in spelling detection and correction in other languages like Bangla etc. Section II contains an introduction of Gurmukhi Script. Section III presents data collection and analysis method used for calculating results.Section IV focuses on difficulties in automatic text error correction in Punjabi Language. Section V describes statistical results found regarding 'Addak' and 'Bindi'.Section VI concludes research work and presents a future scope.

II. INTRODUCTION OF GURMUKHI SCRIPT

The word 'Gurmukhi' literally means from the mouth of the Guru. Gurmukhi script is used primarily for the Punjabi language, which is world's most widely spoken language. Gurmukhi script-consists of 41 consonants called *vianjans*, 9 vowel symbols called *laga* or *matras*, 2 symbols for nasal sounds, one symbol for reduplication of sound of any consonant and three half characters.

Table 1: Gurmukhi Vocabulary

Consonants				
a	A	e		
		s	h	
k	K	g	G	|
c	C		j	J
t	T		f	F
	x			
q	Q	d	D	n
p	P		b	B
	m			

	X r l v V
	S ^ Z z & L
Vowels	w ,i , I , u , U , y , Y, o , O
Semi-Vowels	N , ° , `
Half Characters	*HH R Í*

Vowel Consonants: The consonants of first row (a,A,e) are classified as open syllabics and called vowel consonants or semi consonants or "Matra Vahak" due to their inherent property that they are never used in work without any 'Laga' or 'Vowel'.

Root Consonant: The next two consonants are classified as root class consonants.

There are five such categories namely the Kavarg toli, Chavarg toli, Tavarg toli and the Pavarg toli depending upon the different organs like throat, palate, mouth, tongue and lips, using which they are pronounced or from where they originate.

Antim Group: The last but one group consisting of 5 independent consonants (X,r,l,v,V) is called the "Antim" group

Naveen Group: "Naveen" group (S,^,Z,z,&,L) is the last group which has been introduced to accommodate the words of Persian, Arabic and Sanskrit.Punjabi has three diacritics namely bindi (ਅਾਂ), tippi (ਅੰ) and addak (ਅੱ) used with vowels. These diacritics quite important as their use change the meaning of the words. This paper mainly focuses on addak and bindi.

Addak (`)

This symbol creates the impression of increasing or emphasizing the intensity of letter that follows it, by leaving a small time. In the middle of the word ਬਿੱਲੀ, you can see a small, 'u'-like structure, above the joining line. This is called an addak (ਅੱਧਕ).

Bindi(.) : the bindi is reduced in size further until it has just become a dot (which is what "bindi" actually means). e.g. ਅੈਂ, ਕਾਂ, ਓਂ.

III. DATA COLLECTION AND ANALYSIS

Statistical data for the results was collected from Typing Colleges, Professional typists and Government institutions and private printing presses and every document was carefully scrutinized and the misspelled words were manually collected and analyzed. Out of Text containing more than eight lakh words around 20000 misspellings were found. Different type of analysis has been performed on the corpus like word length analysis, Special character analysis First position error analysis etc.

IV. DIFFICULTIES IN AUTOMATIC TEXT ERROR CORRECTION IN PUNJABI (10)

Though considerable work has been done on automatic spell checking and correction in English language, for Indian language error correction, it has shown more difficulties than that of English because of Indian Language characteristics. The key reasons for difficulties in automatic text error correction in Punjabi are listed below:

(1) Multiple ways of writing the same word.
(2) Difference between Phonetic utterance and the spelling of that word.
(3) Problems regarding the characters of Naveen group elements.
(4) Borrowed words from other Languages (English etc).
(5) Phonetically Similar Character Errors.

V. STATISTICAL RESULT ABOUT '*Addak*' and '*Bindi*'

These characters play an important role in insertion and deletion errors. It is estimated that 37.98% of deletion errors and 30.05% of insertion errors are due to these characters (fig1, fig2,fig3). , For example gl→g`l,cl→c`l,iv`c→ivc, etc.

A. Occurrence of Real word Errors due to Insertion of Addak and Bindi:

Insertion error occurs when at least one extra character is inserted in the desired word. For example gl→g`l,cl→ c`l, here `` is the extra inserted character. These errors also give rise to real word errors .In the above example g`l, c`l are two valid words but they are not the desired word. It is seen that the characters N ,` characters are mostly extra inserted characters. The percentage of insertion N is 17.53 % and the percentage of ` is 12.52% and these two characters contribute around 30% of Insertion errors (fig 1).

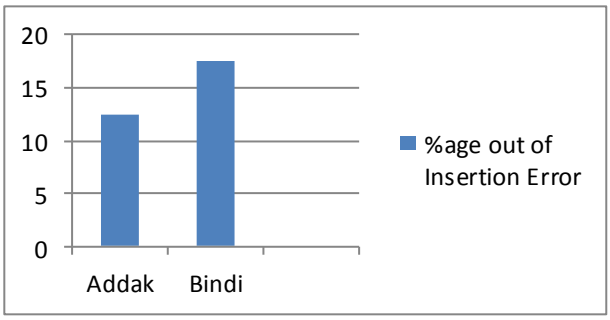

Fig1: Percentage of Addak and Bindi out of total no. of Insertion errors

B. Occurrence of Real word Errors due to deletion of Addak and Bindi:

Deletion error occurs when at least one character is deleted in the desired word, for example g`l →gl ,c`l→ cl.These errors also give rise to real word errors ,for example in the following example
Pu`l→ P`l ,P`l is a valid word but it is not the desired word. It is observed that deletion related errors contribute significantly after substitution errors. It is seen that the characters ` ,N are most commonly missing characters. The percentage of missing ` is 20.51 %and the percentage of missing N is 17.47% and these two characters along contribute to 38% of deletion errors (fig. 2).

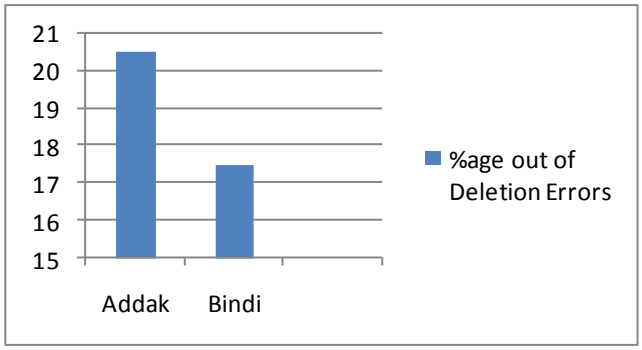

Fig 2: Percentage of Addak and Bindi out of total no. of deletion errors

C. Comparative Graph:

It is estimated that 37.98% of deletion errors and 30.05% of insertion errors are due to these characters (fig3).

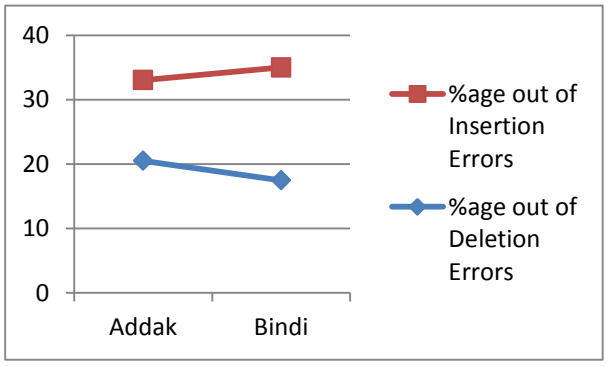

Fig 3: Comparison Graph of percentage of Addak and Bindi in Insertion and deletion errors

VI. CONCLUSIONS

A detailed study has been made on the different type of error analysis of Punjabi Typed text regarding automatic text error correction in Punjabi. This analysis is helpful in creating suggestion list for Punjabi spellchecker. I have done analysis based on, positional effects, first position error analysis, phonetic effects, word length effects etc. This paper mainly focuses on the contribution of 'Addak' and 'Bindi' in Statistical Error analysis of Punjabi typed text. This work can be enhanced for handwritten text pattern analysis and OCR generated error analysis. Analysis can be extended to more number of misspellings and different types of data materials also.

REFERENCES

[1] P. Kundu and B.B. Chaudhuri (1999) *"Error Pattern in Bangla Text"*. International Journal of Dravidian Linguistics. 28(2): pp 49-88.

[2]Pollock, J. J., and Zamora, A. 1983. *Collection and Characterization of spelling errors in scientific and scholarly text.* J. Amer. Soc. Inf. Sci. 34, 1, pp 51-58.

[3] K.W. Church and W.A. Gale (1991) "*Probability scoring for spelling correction"*. Statistical Computing. 1(1): pp 93-103.

[4] K. Kukich (1992) *"Techniques for automatically correcting words in text"*. ACM Computing Surveys. 24(4): pp 377-439.

[5] F.J. Damerau (1964) "A technique for computer detection and correction of spelling errors".Commun. ACM. 7(3): pp 171-176.

[6] Morris, Robert & Cherry, Lorinda L, *'Computer detection of typographical errors'*, IEEE Trans Professional Communication, vol. PC-18, no.1, pp54-64, March 1975.

[7] Yannakoudakis, E. J., and Fawthrop, D. 1983a. *"An* intelligent spelling corrector". Inf. Process. Manage. 19, 12, 101-108.

[8] Yannakoudakis, E. J., and Fawthrop, D. 1983b. *"The rules of spelling errors".* Inf. Process. Manage. 19, 2, 87 99.

[9] Wagner, Robert A. & Fischer, Michael J, *'The string-to-string correction problem'*, Journal of the A.C.M., vol.21, no.1, pp168-173, January 1974.

[10] R.E. Gorin (1971) *"SPELL: A spelling checking and correction program"*, Online documentation for the DEC-10 computer.

[11] Meenu Bhagat,*"Difficulties in automatic text error correction in Punjabi"*, In Proceedings of "International Conference on Control Communication and Computer Technology" 6-7 th Aug 2011, New Delhi.

Authors Profile

Mrs Meenu Bhagat pursed Bachelor of Computer Science & Engg from Baba Banda singh Bahadur Engg. College,Fatehgarh Sahib,Punjab affiliated with Punjab Technical University in 2001 and done Master of Engineering from Thapar University in year 2003.She is currently pursuing Ph.D. and currently working as Senior Assistant Professor in Punjab University Swami Sarvanand Giri Regional Centre,Hoshiarpur,Punjab, Department of Computer Science and Engineering since 2006. She is a life member of the ISTE since 2005. She has published more than 10 research papers in international conferences and journals. Her main research work focuses on Natural Language Processing,Software engineering,spell checking Interfaces etc. She has more than 12 years of teaching experience .

