ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 7, July 2016

Statistical Review of First Position Errors in Punjabi Typed Text

Meenu Bhagat
Assistant Professor,
Department of Computer Science & Engineering
Punjab University SSG Regional Centre Hoshiarpur,
Punjab, India
meenubhagat@yahoo.com

***Abstract*:--Positional analysis of any language is useful in spellchecker, Natural Language Interfaces, OCR and language related technology development etc .Though considerable work has been done in the area for English and related languages, the Indian Language scenario is still far behind.. This paper focuses on the role of First position errors in Non-word Error Distribution of Punjabi Typed Text. This paper is based on the analysis done on 20000 misspelled words generated by typists.**

***Keywords*:-- Phonetic, Kavarg, Naveen, Gurmukhi, Kavarg.**

I. INTRODUCTION

Kukich[1] has discussed the different techniques for automatically detection and correction of misspellings and the identify the various factors affecting the spelling errors patterns of words in English. Damerau [2] worked on a technique for computer detection and correction of spelling errors in English language. Church and Gale [3] have done a probability scoring of spelling correction. Chaudhuri and Kundu [4] have done an elaborative analysis on error pattern generated by Bangla text patterns and made a reversed word dictionary and phonetically similar word grouping based Bangla spellchecker.

Error can be of two types, namely, Non-word error and Real-word error.A Real word error occurs when a word is misspelled as another valid word which is not proper for the context.An instance of a real word error is PU`L→ P`L,here the valid Punjabi word PU`L is misspelled as another valid word P`L . Though the word is syntactically correct,It is semantically incorrect within the context of sentence.

If a string of characters is separated by spaces or punctuation marks it is called a Candidate string. A Candidate string is said to be valid word if it has a meaning otherwise, it is a non-word. In each case the problem is to detect the Error and suggest correct alternatives or automatically replace it with correct word.

Pollock and Zamora [5] 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 technique. 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 to introduce the concept of applying dynamic programming techniques to the spelling correction problem to increase computational efficiency.

A “reverse” minimum edit distance technique was used by Gorin [10] in the DEC-10 spelling corrector and by Durham et al.[11] in their command language corrector. Church and Gale [12] and Kernighan et al [13] also used a reverse technique to generate candidates for their probabilistic spelling corrector.

II. BRIEF OVERVIEW OF GURMUKHI SCRIPT (14)

Gurmukhi script is used primarily for the Punjabi language, which is world's most widely spoken language. The word 'Gurmukhi' literally means from the mouth of the Guru. Gurmukhi script is syllabic in nature. Gurmukhi script-consists of 41 consonants called vianjans, 2 symbols

ISSN (Online) 2394-2320

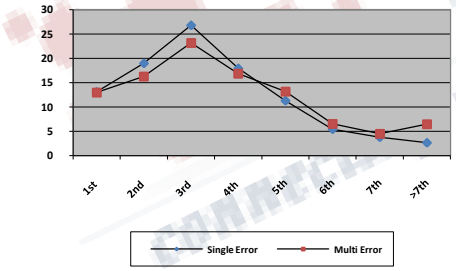
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 7, July 2016

for nasal sounds, 9 vowel symbols called laga or matras,one symbol for reduplication of sound of any consonant and three half characters.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'. The next two consonants are classified as root class consonants. The rest of the consonants except to the last two groups namely the - "Antim" and "Naveen" group, are categorized according to their phonetic structure.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.The last but one group consisting of 5 independent consonants (X,r,l,v,V) is called the "Antim" group and the last group is the (S,^,Z,z,&,L). "Naveen" group has been introduced to accommodate the words of Persian, Sanskrit and Arabic.

III. POSITIONAL ANALYSIS

The positional analysis plays an important and significant factor in the error pattern study. This can lead us to error zone of high probability. It has been found out that patterns for the positional mistakes is almost similar in both single/multi-error misspellings. The maximum of the mistakes occur at the third position and the error zone decreases after 3rd position.

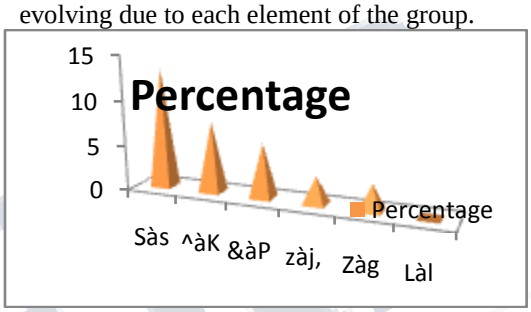
Figure 1 Position wise distribution of misspellings



Pollock and Zamora[5] found that 3.3% of the 50000 misspellings involved first letter and **Yannakoudakis and Fawthrop[7-8]** observed a first position error rate of 1.4% in 568 typing errors.In Punjabi language, first position error rate is higher than english language i.e.13% in Punjabi as compared to English.About 75.49% errors lie in first four positions. It is observed that in single error misspellings 13.10% and 13.0% in multi error misspellings are found to be first position errors.

This rate is more than as expected. Concluded reasons are:

I. Out of the total first position misspellings, 32.91% were the misspellings who have mistakes due to (S,^,Z,z,&,L)i.e. where the typist has typed S→s,^→K,Z→g,z→j,&→P,L→l. It means at least 32.91% of the first position misspellings are due to substitution errors. Though there are many more other substitution pairs (for example n→l, a→v) that are also found. It is clearly signifying the probability of the substitution error at the first position. Fig 2 is showing the distribution of errors evolving due to each element of the group.



Shifted and Unshifted modes of typing: For example n→l,a→ v.
Multiple forms for same word are found in Punjabi Language ,for example jyhw →ijhw, vIcwr→ivcwr .The percentage of Substitution of the above word pairs for out of the total no. of first position error misspellings is 3.15%.

IV. CONCLUSION

A detailed study has been made on the first position error analysis of Punjabi Typed text. This analysis is based on the detailed analysis based on the positional analysis that can be helpful in creating suggestion list of Punjabi spellchecker. In addition to this various other effects like phonetic effects, word length effects has also been studied. Besides the usual typing mistakes, the other reasons for first position errors in Punjabi language are due to:

1. Due to Naveen group substitution errors like S→s,^→K,Z→g,z→j,&→P,L→l .
2. Shifted and Unshifted modes of typing.
3. Due to non-standardization of Punjabi spellings.
4. Due to phonetic similarities of various consonants and vowels.

REFERENCES

[1] K. Kukich (1992) "Techniques for Automatically Correcting words in Text". ACM Computing Surveys. 24(4): 377-439.

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

[3] K.W. CHURCH AND W.A. GALE (1991) "probability scoring for spelling correction". statistical computing. 1(1): 93-103.

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

[5] 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, 51-58.

[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. & Fawthrop, D, 'An intelligent spelling error corrector', Information Processing and Management, vol.19, no.2, pp101-108, 1983. (1983b)

[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] Durham, I, Lamb, D.A, & Saxe, J.B, 'Spelling correction in user interfaces', Communications of the A.C.M., vol.26, no.10, pp764-773, October 1983.

[12] Gale and Church, 1991[b] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. In Proceedings of the 29th Meeting of the ACL, pages 177-184. Association for Computational Linguistics, 1991.

[13] M.D. Kernighan, K.W. Church, and W.A. Gale. 1990. A spelling correction program based on a noisy channel model. In Proceedings of the Thirteenth International Conference on Computational Linguistics, pages 205-210.

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