International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072

Single/Multi-Error Misspellings in Punjabi Typed Text

Meenu Bhagat

Assistant Professor, Department of Computer Science & Engg., Punjab University

SSG Regional Centre Hoshiarpur, Punjab, India.

Abstract—Statistical analysis of spelling error patterns in a language plays an important role in automatic spelling error detection and correction. This research study identifies and analyses the spelling error trends in Punjabi. The statistical analysis of error trends is based on a real time data collected from different sources .Both traditional (insertion, deletion, transposition, substitution, run-on, split word error)) and language specific error trends positional analysis, word length effects, phonetic errors, first position error analysis, keyboard effects etc. are identified and analyzed. This paper focuses on the contribution of Single/Multi-Error misspellings in Punjabi Typed Text. It also discusses previous analysis results about spelling error patterns found in other languages and offers new insights on them. This paper is based on the analysis done on 20000 misspelled words generated by typists.

Keywords: Multi-error, Real word, Gurmukhi, Non-word.

I. Introduction

This paper reports results on a research on Single/Multi-error misspellings found in Punjabi Language. Error can be of two types, namely Non-word error and Real-word error. 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. Real Word error is a valid word but not the intended word in the sentence, making the sentence syntactically or semantically incorrect. In each case the problem is to detect the Error and suggest correct alternatives or automatically replace it with correct word.

Kukich[1] has discussed the different techniques for automatically detection and correction of misspellings and identification of 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.

Pollock and Zamora [5] aimed at discovering probabilistic tendencies, such as which letters and position within a word are most frequently involved in errors, to devise similarity key based technique. Morris and cherry [6] worked on devising 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] introduced the concept of applying dynamic programming techniques to the spelling correction problem to increase computational efficiency.

Gorin [10] and Durham et al. [11] used “reverse” minimum edit distance technique in the DEC-10 spelling corrector and r command language corrector respectively. Reverse technique was also used by Church and Gale [12] and Kernighan et al [13] to generate candidates for their probabilistic spelling corrector. The statistical data we provide on spelling error patterns in Punjabi and their comparison with other data in other languages are the novel contribution of this paper.

II. Single/Multi-Error Distribution

Single error misspellings are the misspellings in which a word contains one error where as multi-error misspellings where a single word contains multiple instances of errors. Kukich [1] has found upwards of 80% misspellings to be single error misspellings and most misspellings tend to be within two characters in length of the correct misspelling. Damerau[2] found that approximately 80% of all misspelled words contained a single instance of one of the following four

types of errors: **insertion**, **deletion**, **substitution** and **transposition**. An analysis was also carried for different type of Single/Multi-error misspellings for Punjabi typed Text and It has been found that out of the total no. of misspellings, 91.13% were the single error misspellings and 8.87% were multi error misspellings. While for English language **Pollock and Zamora[5]** found that only 6% of 50000 nonword spelling errors in the machine readable databases they studied were multierror misspellings and Coversely , **Mitton[15]** found that 31% of the misspellings in his 17001 word corpus of handwritten essays contained multiple errors. It is observed that majority of the multi-error misspellings contain two mistakes (see Fig 1).

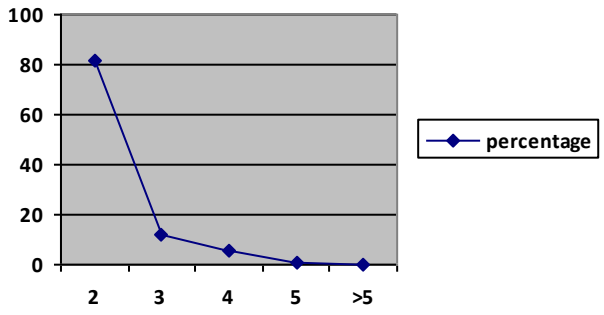

Fig 1 Showing the Percentages of no. of mistakes in a word

III. First Position Errors

The percentage of first position errors in Punjabi language is considerable. It is observed that in single error misspellings 13.10% and 13.0% in multi error misspellings are found to be first position errors.

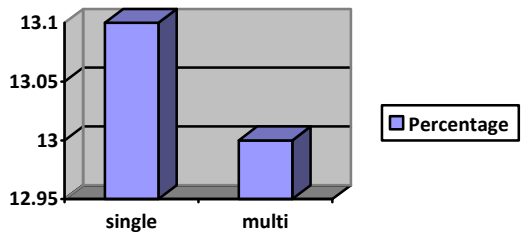

Fig 2 Showing the Percentages single/multi-error misspellings

IV. Transposition Error Analysis in Single/Multi Error Misspellings

Transposition error occurs when two adjacent characters of a word are typed in swapped manner, for example rwq→ rqw. In the above word w →q are transposed character pairs.

It is found that these transposition errors (like substitution) also give rise to real word errors ,for example

krm →kmrw whereas kmr is a valid word. The Percentage of transposition errors is 1.85% and 1.43% in single and multi-error misspellings respectively. No prominent transposition character pairs were found.

V. Positional Analysis on Single/Multi-error Misspellings in Punjabi Language

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 are 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.

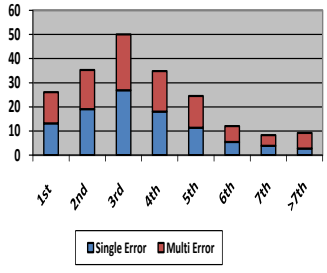

Figure 3 Position wise distribution of Single/Multi-error misspellings

It is observed that in single error misspellings 13.10% and 13.0% in multi error misspellings are found to be first position errors.

VI. Conclusion

This paper reports findings from the elaboration of a statistical analysis of spelling errors for Punjabi Typed Text. It also discusses previous generalizations about spelling error patterns found in other languages and offers new insights on them. A detailed study has been made on the single/multi error misspellings of Punjabi Typed text. This analysis is based on the detailed analysis of different types of errors. This type of data provides clues about which error patterns should be promoted to enhance the generation of suggestions lists for corrections in Punjabi spellchecker. In addition to this various other effects like phonetic effects, word length effects has also been studied.

VII. 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.

[15] Roger Mitton (1987),"Spelling checkers, spelling correctors and the misspellings of poor spellers", Information Processing and Management: an International Journal, v.23, and pp. 495-505.