A Comparative Study of Various Errors Pattern for Different Languages and its types

Rajesh Sharma
Department of Computer Applications
Assistant. Professor. CT Institution of Engineering Management and Technology, Jalandhar, India

Gurpreet Singh Lehal
Department of Computer Science
Professor, Punjabi University, Patiala, India

Abstract- Natural Language Processing (NLP) is a field of Computer Science. It is concerned with the interaction with computer and human languages. Modern NLP algorithms are based on Machine learning. In advanced computer system, text errors are to be expected in various forms such as OCR errors, typing errors, query based search engine errors, translation and transliterate errors etc. These errors present a grave confront to downstream processes that attempt to make use of such data. The capability to accurately recognize characters are enormously essential for various forms of automated data processing and has wide application. Since the dawn of the computing era, information has been represented digitally so that it can be processed by electronic computers. This paper describes the various types of text related computer error in different language. To search out the various errors in Indian language are very challenging.

Keywords: Errors, Typographic, OCR Errors, Translation and Transliterate Error, Query Based search Engine Error.

I. Introduction

The expression Error generally means that "Something is going wrong". The cause of occurrence of errors may be missing letters, extra letters, misspelled letters, merged words, splitted words or disordered letters. Error is categorized in two groups. The earliest is real word error and next is non-real word error. A word that is not present in lexicon is called non-real word and a word that is valid and is present in lexicon is called real word. Detection of non-real word errors are very challenging and hard task of NLP applications.
Errors can be classified into many types.

1.1 Typing Errors

Typing error or Typographic errors are occurred due to the typist by mistake presses the wrong key, presses two keys, and presses the keys in the wrong order.

(a) *Substitution Errors (SE):* In this type, one or more characters are replace by another character and the resulting word may be real or non real. For example

Correct word	**Incorrect Word**	**Description**
ਜਥੇਦਾਰ	ਜਥੇਧਾਰ	ਦ is substituted by ਧ

(b) *Deletion Error (DE):* Similarly in this type, one or more characters are deleted from the real word. For example

Correct word	**Incorrect Word**	**Description**
ਯੂਨੀਵਰਸਿਟੀ	ਯੂਨੀਵਸਿਟੀ	ਰ is deleted

(c) *Insertion Errors (IE):* When one or more extra characters are added in the correct word. For example

Correct word	**Incorrect Word**	**Description**
ਫਿਲਮ	ਫਿਲਾਮ	ਲ is replaced by ਲਾ

(d) *Run-On-Error Or Merged Error (ROE):* Basically these types of errors are raised due to misplaced of white spaces between two or more words. For example

Correct word	**Incorrect Word**	**Description**
ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ	ਪੰਜਾਬੀਯੂਨੀਵਰਸਿਟੀ	No space between ਪੰਜਾਬੀਯੂਨੀਵਰਸਿਟੀ

(e) *Transposition Errors (TE):* This type of error occurs when two adjacent characters are swapped. For example

Correct word	Incorrect Word	Description
ਸਿਹਤ	ਸਿਤਹ	ਹ and ਤ is swapped

(f) *Split Word Errors (SWE):* When extra spaces are additionally added in between in one or more words are called split word error. For Example

Correct word	Incorrect Word	Description
ਹਸਪਤਾਲ	ਹਸ ਪਤਾਲ	Extra space is added in between ਹਸਪਤਾਲ, word become ਹਸ ਪਤਾਲ

(g) *Visual Error (VE):* When a word looks visually correct but inside it is incorrect. This type of error typical comes in Punjabi typing. For example

Original Text: ਉਹ ਦੁੱਧ ਪੀਂਦਾ ਪਿਆ ਹੈਂ।

Output Text: ਉਹ ਦੁੱਧ ਪੀਂਦਾਂ ਪਿਆ ਹੈਂ।

In above example ਪੀਂਦਾ visually looks to be rightly spelled but it is stored as ਪੀਂਦਾਂ

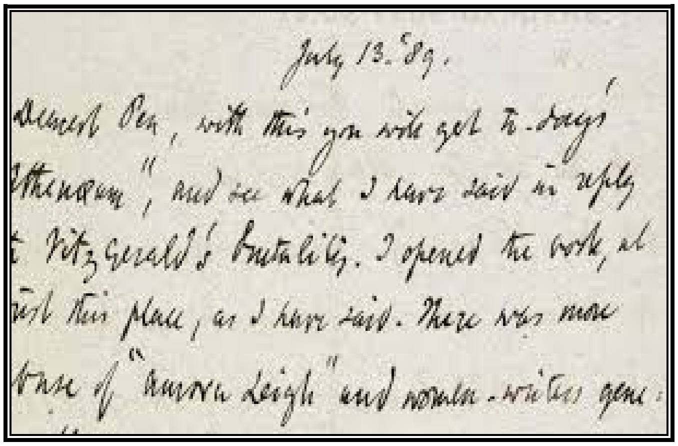
(h) *Phonetically Similar Character Errors:* Phonetic error are those error which the writer substitutes a phonetically correct but orthographically incorrect sequence of characters for the required word.

Class 1		***Class 2***		***Class 3***		***Class 4***	
ਜ	ਝ	ਸ	ਸ਼	ਰ	◌੍ਰ	◌ੁ,	◌ੂ
ਬ	ਭ	ਖ	ਖ਼	ਹ	◌੍ਹ	◌ੋ	◌ੋ
ਠ	ਣ	ਗ	ਗ਼	ਵ	◌੍ਵ	◌ੌ	◌ੋ
ਗ	ਘ	ਲ	ਲ਼			◌ਂ	◌ੰ
		ਫ	ਫ਼				

1.2 OCR Errors

Optical Character Recognition (OCR) is the process of transforming images of handwritten or typewritten text into machine-editable text **(Cheriet, Kharma, Liu, and Suen, 2007).** There are often mistakes in the scanned texts as the OCR system occasionally misrecognizes letters and falsely identifies scanned text, leading to misspellings and linguistic errors in the output text **(Niklas, 2010**). OCR errors can be divided into following types:

(a) *Old-style language:* Some old letters can be used in these kinds of texts such as the "long I" in old English writing which is often confused with the lower-cased " I " in various Roman typefaces and in blackletter. For example the word is "you write get to" can be recognised by automatic systems as "you wide get to".



July 13. '89.
Dearest Pen, with this you will get to-day's
"Athenaeum", and see what I have said in reply
to Fitzgerald's brutality. I opened the work, at
just this place, as I have said. There was more
... of "Aurora Leigh" and woman-writers gene-

Figure 1: Old Style Writing Text

(b) *Punctuation errors:* especially when we have a poor scanning quality, punctuation character misrecognition can cause commas or full stops to often occur in the wrong positions.

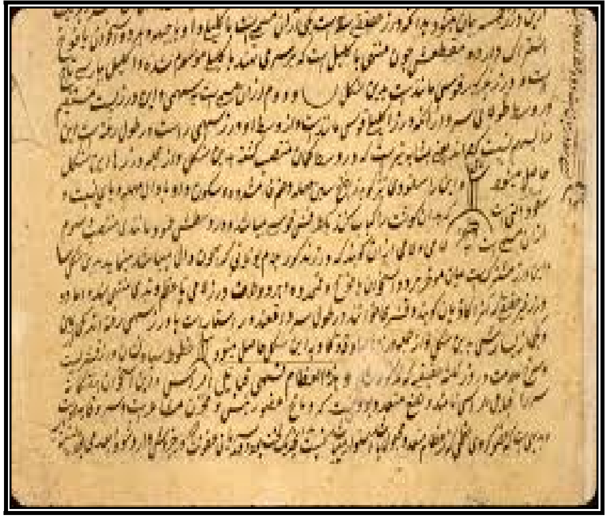


Figure 2: Poor quality Scanning

(c) *Character format:* variations in font can also prevent accurate character recognition which leads to wrong word recognition (e.g. ”iii" instead of "m"). It can also cause the case-sensitivity errors when lower and upper case characters can be mixed up. Another example like ਪ insead of ਥ and in shahmukhi like ب (ਥ) instead of پ (ਪ)

(d) *Character Insertion, Deletion and Substitution (IDS):* It is often the case that one or more characters are substituted or deleted, or that a character is wrongly inserted in the middle of a word like ਉਹ ਦਿੱਲੀ ਜਾ ਰਿਹਾ ਹੈ instead of ਊਹਦਿੱਲੀ ਜ1 ਰਿਗ ਹੈ

(e) *Segmentation errors:* Unusual spacings in lines, words or characters lead to misrecognitions of white-spaces in some cases which can cause segmentation errors (e.g."ਉਡਤਾ ਪੰਜਾਬ " instead of "ਉਡ ਤਾਪੰ ਜਾਬ").

(f) *Word meaning:* some misrecognized characters can generate new words which are often wrong in context but correctly spelled (e.g."ਓਹ ਜਾਣਾ ਚਾਵੇ ਤਾ ਜਾਵੇ" instead of " ਓਹ ਜਾਣਾ ਚਾਵੇ ਤਾ ਪਾਵੇ").

1.3 Query Refinement in Search Engine Error

Query refinement involves reformulating *ill-formed* search queries in order to enhance relevance of search result. Query refinement typically includes a number of tasks such as spelling error correction, word splitting, word merging, phrase segmentation, word stemming and acronym expansion.

Any set of query keywords contains a large amount of noise data, such as words in foreign languages or misspelled words **(Alfonseca et al., 2008).** It typically includes a number of tasks such as spelling error, word splitting, Word Merging, Phrase Segmentation, Word Stemming, acronym expansion. Query refinement is by nature a structured prediction problem which seeks to predict the latent structure of an observation sequence. **(Jiafeng et al., 2008)**.

(a) Spelling Errors: Spelling is defined as writing a word or group of words to form a specific leeter in particular order.But spelling errors refer to write a wrong word which is caused by wrong typing of a word.Like ਜਸਪ੍ਰੀਤ is a right word but while typing it is written as ਜੁਸਪ੍ਰੀਤ.Four types of operation are used, that are

- **Deletion** which includes missing of letters like ਜਪ੍ਰੀਤ is a wrong word.
- **Insertion** which include addition of unwanted letter like ਜਸਹਪ੍ਰੀਤ.
- **Substitution** which includes replacement of one letter with other like ਮ is replaced with ਸ word become ਜਮਪ੍ਰੀਤ.
- **Transportation** involves swapping of letters like ਮਨਮੋਹਨ here ਹ and ਨare swapped ਮਨਮੋਨਹ

(b) Word Splitting Errors: refer to splitting the one word into words. Like the word ਚਿੜੀਆਘਰ then the space is added after ਚਿੜੀਆ. So the word the splitted into ਚਿੜੀਆ and ਘਰ both ਚਿੜੀਆ, ਘਰ are real words.

(c) Word Merging Errors: are the errors which are caused due to merging of one or two words after merging they form a different sense like the original words are ਅਸੀਂ ਦਿੱਲੀ ਗਏ but after merging it becomes ਅਸੀਂਦਿੱਲੀਗਏ

(d) Phrase Segmentation Errors: basically phrase is group of words which is standard and is used. Phrase segmentation errors refer to the segmenting the phrase into wrong one like ਸਤਿ ਸ੍ਰੀ ਆਕਾਲ is a phrase but after segmenting error it becomes ਸਤਿ ਸ੍ਰੀਆ ਕ1ਲ.

(e) Word Stemming: is the formation of words from one base word. It is explained by basis word ਢੋਲits stemming words areਢੋਲੀ, ਢੋਲਿਆਂ, ਢੋਲਕੀ etc.

(f) Acronym expansion Errors: Acronym expansion functions for search are considered an accessibility feature that is useful to people who have difficulties typing like ਭਾਰਤੀ ਜਨਤਾ ਪਾਰਟੀ make be expanded or pronounced as ਬੀਜੇਪੀ, ਬੀ.ਜੇ.ਪੀ., ਭਾਜਪਾ, ਭਾ.ਜਾ.ਪਾ.

1.4 Translation/Transliteration Errors:-

A Translation/Transliteration would induce number of errors. These errors are of four broad types, namely

(a) Inflection Error: Inflections carry important grammatical information about the structure and meaning of words, can change the interpretation of a word (and sentence), and hence are important for both meaning and grammar.(For Example हम निरक्षरों की श्रेणी में आ जायें after translate/transliterate it becomes □□□ □□□□□□□□ □□ □□□□□□ □□□□ □ □□□□)

(b) Non Technical Ambiguity: If you say that there is ambiguity in something, you mean that it is unclear or confusing, or it can be understood in more than one way. Translator could not decide which word was more appropriate (For Example:- पीढ़ी के कम्प्यूटरों का आकार बड़ा होता after translate/transliterate it becomes □□□□□□ □□ □□□□□□□□□ □□ □□□□ □□□□ □□□□□ and ऑफ लाइन it becomes □□ □□□□)

(c) Translation Error: Translation error can be caused by misunderstanding of the translation brief or of the content of the Source translation (ST), by rendering the meaning of the source translation(ST) accurately, by factual mistake, terminological or stylistic flaws and by different kinds of interferences between ST and TT. (For example:- इस प्रकार संग्रहीत प्रोग्राम के सिद्धान्त का जन्म हुआ after translate/transliterate it becomes □□ □□□□□□ □□□□□□□ □□□□□□□□ □□ □□□□□□ □□ □□□ □□□□)

(d) Word out of Vocabulary: Out-Of-Vocabulary words are unknown words that appear in the source translation but not in the recognition vocabulary.(For example:- कई दशक पहले आविष्कारित ये कम्प्यूटर after translate/transliterate it becomes □□ □□□ □□□□□□ □□□□□□□□□ □□ □□□□□□□)

1.5 Miscellaneous Errors:-

(a) Different forms of writing the same word: In Indian language especially in Punjabi and Urdu there are various ways to write the same things in different ways and all these ways may be accurate. In Punjabi language there is no standardization of spellings of various words and the same word has been written in

different forms with various typesets. e.g. ਜਲੰਧਰ, ਜਾਲੰਧਰ, ਜੱਲ਼ੰਧਰ. Here, all the three words are same meaning but have different iterations of the same word. So it is affecting the estimated structure and volume of the dictionary.

(b) *Multiple character with same Sound:* In Punjabi and Urdu script, there are many character sound whose sound has more than one latter. This is the major reason for spelling mistakes. For example ਭ and ਬ, ਠ and ਤ etc.

(c) *Difference between spelling and Diacritics utterance***:** In Indian languages, there are lot of words whose accent is dissimilar from its spelling. e.g.

ਸ਼ੈਹਰ	ਸ਼ਹਿਰ
ਔਂਦਾ	ਆਉਂਦਾ
ਪਾਅਣੀ	ਪਾਣੀ
ਅਧਾਰ	ਆਧਾਰ

(d) *Problems due to Nasal Sound, Naveen and Nukta character symbols***:** These can be occurred due to mistakenly presence and absence of Naveen Group characters or Nukta symbols in Punjabi and Nasal Sounds for Urdu . For Example:-

ਸ	ਸ਼
ਖ	ਖ਼
ਗ	ਗ਼

(e) *Words borrowed from other Languages***:** The size of lexicon is increasing day by day as new words borrowed from other languages and therefore there is no limit of size of lexicons. Rich Indian Languages have included many words borrowed from other languages especially from English, Urdu. E.g. ਸਟੇਸ਼ਨ, ਬਸਸਟੈਂਡ-, ਪਲੇਟਫਾਰਮ, ਗਜ਼ਲ, ਗੁਸਲਖਾਨਾ, ਮੁਬਾਰਕਾਂ

II. CONCLUSION

In this paper, we have discussed about outline of various error pattern in various local languages. Our analysis of these error have based on above mentioned error patterns. In addition the usual mistakes are due to: Multiple forms of the same word, Slight difference between the pronunciation and spellings of some of the common words, Phonetic similarities of various consonants and vowels, Borrowed words from other languages, Unnecessary Insertion, deletion, substitution etc. In future, we will study various error detect and correction techniques and try to develop algorithm to detect and correct these kind of various errors.

REFERENCES

[1] Kazem Taghva, Eric Stofsky "OCRSpell: an interactive spelling correction system for OCR errors in text" International Journal on Document Analysis and Recognition" Springer-Verlag 2001, pp 125-137.

[2] Jiafeng Guo, Gu Xu, Hang Li, Xueqi Cheng "A Unified and Discriminative Model for Query Refinement" SIGIR'08, July 20–24, 2008, Singapore, pp 379-386.

[3] Eric Brill, Silviu Cucerzan "Spelling correction as an iterative process that exploits the collective knowledge of web users" Conference on Empirical Methods in Natural Language Processing, 2004, pp 293-30.

[4] Serena Jeblee , Houda Bouamor, Wajdi Zaghouani and Kemal Oflazer "An SMT-based System for Automatic Arabic Error Correction" Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), October 25, 2014, Doha, Qatar. pp 137-142.

[5] Huizhong Duan, Bo-June (Paul) Hsu "Online Spelling Correction for Query Completion". International World Wide Web Conference Committee (IW3C2), March 28–April 1, 2011, Hyderabad, India

[6] Youssef Bassil, Mohammad Alwani "OCR Context-Sensitive Error Correction Based on Google Web 1T 5-Gram Data Set" American Journal of Scientific Research 2012, ISSN 1450-223X Issue 50 , pp 14-25.

[7] Karen Kukich "Techniques for Automatically Correcting Words in Text" ACM Computing Surveys, Vol. 24, No. 4, December 1992.

[8] Jiang-Ming Yang, Rui Cai, Feng Jing, Shuo Wang, Lei Zhang and Wei-Ying Ma, "Search-based Query Suggestion", CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

[9] Hema P. H., Sunitha C "Spell Checker for Non Word Error Detection: Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March-2015, IJARCSSE, pp 360-363.

[10] Lehal G S, Chandan Singh, "A Gurmukhi script recognition system". In Proceedings 15th International Conference on Pattern Recognition, Barcelona, Spain 2000, vol 2, pp 557–560.

[11] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen, *Character Recognition Systems: A Guide for Students and Practitioners*, Wiley-Interscience Publication, (2007).

[12] Kai Niklas. Unsupervised Post-Correction of OCR Errors PhD thesis, Leibniz University Hannover, in Germany 2010.