ijecs open access www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 9 September 2014 Page No. 8030-8032

Spell Checking and Error Correcting System for text paragraphs written in Punjabi Language using Hybrid approach

Amanjot Kaur,* Dr. Paramjeet Singh, Dr. Shaveta Rani

M.tech Student CSE Department
GZS PTU Campus Bathinda
amanjot_109@yahoo.co.in
Assistant Professor, CSE Department
GZS PTU Campus Bathinda
param2009@yahoo.com
Assistant Professor, CSE Department
GZS PTU Campus Bathinda
garg_shavy@yah00.com

Abstract---The spell checker is the basic necessity for composing any documentation in any language. Spell checker is software that analysis the incorrect word and provide their most possible correct word. Make a spell checker in Indian language is a very challenging and uphill task. Punjabi is world's 14th mostly used language. Work done in Punjab language is very challenging task. In a computer system a Punjabi words is typed in different manners because Punjabi language has more than 40 different fonts. Punjabi is mother tongue for more than 110 million people in the world. This paper describes the techniques used in a spell checker. Natural language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. Modern NLP algorithms are based on machine learning, especially statistical machine learning.

Keywords-- Spell checker, Punjabi, Error, NLP.

I. INTRODUCTION

A spell checker is a technique which identifies the incorrect or misspelled words and replaces them with the best possible combination of correct words. For find incorrect word firstly system checks the word in the dictionary. If the word is finding in the database then it assume to be correct word and if it is not present in the database then system assume this word incorrect and perform the required process to generate best possible combination of correct word. This paper describes the various techniques used for a spell checker such as Rule based technique, dictionary lookup approach and N-Gram technique.

II. Types of Errors

Spelling and typing errors are common in documentation made by human. The problem of detecting error in words and automatically correcting them is a great research challenge. The word error can be divided in two types i.e., non-word error and real-word error. Errors may be of missing letters, extra letters, misspelled letters, or disordered letters. Some of the common errors in a text are as follows:-

ਗਲਤ ਸ਼ਬਦ	ਸਹੀ ਸ਼ਬਦ	ਗਲਤ ਸ਼ਬਦ ਦੇ ਕਾਰਣ
ਅੰਖ	ਅੱਖ	“ ਂ” ਗਲਤ ਹੈ
ਸ਼ੁਭ	ਸ਼ੁੱਭ	“ ੁ”, “ ੱ” ਲੱਗਣਾ ਹੈ
	ਸ਼ੁੱਧ	“ ੁ”, “ ੱ” ਅਤੇ “ ਧ” ਲੱਗਣਾ ਹੈ
ਹਲਟਲ	ਹਲਚਲ	“ਟ” ਦੀ ਜਗ੍ਹਾ“ਚ” ਲੱੱਗਣਾ ਹੈ
ਠਮਲਾ	ਗਮਲਾ	“ਠ” ਦੀ “ਗ” ਲੱਗਣਾ ਹੈ
	ਕਮਲਾ	“ਠ” ਦੀ ਜਗ੍ਹਾ “ਕ” ਲੱਗਣਾ ਹੈ

Since their inception, computers have been exploited broadly to solve and automate complex problems related to diverse domains and fields including mathematics, sciences, education, medicine, gaming, multimedia, and linguistics. In effect, computational linguistics also known as natural language processing (NLP) is a field of both computer science and linguistics that deals with the analysis and processing of human languages using digital computers. NLP has also many applications, they include but not limited to Automatic Summarization, Machine Translation, Part-of-Speech Tagging (POS), Speech Recognition (ASR), Optical Character Recognition (OCR), and Information Retrieval (IR). Spell-checking is yet another significant application of computational linguistics whose research extends back to the early seventies when Ralph Gorin built the first spell-checker for the DEC PDP-10 mainframe computer at Stanford University. By definition, a spell-checker is a computer program that detects and often corrects misspelled words in a text document. It can be a standalone application or an add-on module integrated into an existing program such as a word processor or search engine. Fundamentally, a spell-checker is made out of three components: An error detector that detects misspelled words, a candidate spellings generator that provides spelling suggestions for the detected errors, and an error corrector that chooses the best correction out of the list of candidate spellings. All these three basic components are usually connected underneath to an internal dictionary of words that they use to validate and look-up words present in the text to be spell-checked. However, as human languages are complex and contain countless words and terms, as well as domain-specific idioms, proper names, technical terminologies, and special jargons, regular dictionaries are insufficient to cover all words in the vocabulary of the language. A problem formally known as OOV short for Out of Vocabulary or Data Sparseness which regularly leads to false-positive and false-negative detection of out-of-dictionary words.

III. EXISTING WORK

There are various techniques available for detection and correction of the spellings of Punjabi language which are discussed as follows:

Rule-based Techniques

Rule-based methods are interesting approach used in spell checking. In this the system works by having a collection of rules that capture common spelling and typographic errors and applying these rules to the misspelled word. Ostensibly these rules are “converses” of common errors. Each correct word generated by this process is taken as a correction suggestion. For example:-

ਗਲਤ ਸ਼ਬਦ	ਸਹੀ ਸ਼ਬਦ	ਨਿਯਮ
ਅੀਦਤ	ਆਦਤ	ਅ ਸ਼ਬਦ ਨਾਲ “ੀ” ਨਹੀ ਲੱਗਦੀ
ਊਾਠ	ਊਠ	ਉ ਸ਼ਬਦ ਨਾਲ “ਾ” ਨਹੀ ਲੱਗਦਾ

Edit distance can be viewed as a special case of a rule-based method with limitation on the possible rules

Dictionary Lookup Technique:-

Dictionary-lookup method, which this paper introduces, is a Spellchecked technique. By checking strings of a Punjabi word in the dictionary or the database, it tells a word is correct if it found in the database. If strings not appear in the dictionary or the database, it is the incorrect word. Because of the high precision, Dictionary-lookup method is considered the most important error detection techniques.

N-gram approach:-

N-gram analysis is defined as a process to detect wrong spelled words in a document. Rather of comparing every word in a dictionary, n-grams are used. If an empty or deficient n-gram is found, the word is assumed as an incorrect, otherwise it assume to be correct. An n-gram is a collection of ensuing characters of length N. If N is 1 then the term used is a unigram, if N is 2 then the term is a Bigram, if N is 3 then the term is trigram and so on. Each string that is involved in the comparison process is divided into pair of adjacent N-grams. The n-grams algorithm is also referred as "language independent" or a "neutral string matching algorithm"

IV. CONCLUSIONS

In this paper we have surveyed the area of Spell checking techniques. We have discussed numerous detection and correction methods that are helpful in finding the errors. We have to develop an online Punjabi spell checker system using NLP. The system detects the wrong words and provides their maximum best possible correct words. A large database is to be created for providing better solutions. In future an algorithm that is based on dictionary lookup techniques, rule based techniques and statistical machine translation based techniaque can be developed to improve the results of the system.

V. REFERENCES

1.Gurpreet Singh Lehal(2007), "design and implementation of Punjabi spell checker", International journal of systemic cybernetics and informatics, pp.70-75.

2. Rupinderdeep Kaur and Parteek Bhatia, "Design and Implementation of SUDHAAR-Punjabi Spell Checker," International Journal of Information and Telecommunication Technology, Vol. 1, Issue 15 May, 2010.

3. Neha Gupta &PratisthaMathur,"*Spell Checking Techniques in NLP: A Survey*," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 12, December 2012.

4. Youssef Bassil& Mohammad Alwani, "Context-sensitive Spelling Correction using Google Web IT 5-Gram Information," Department of Computer and Information Science, Vol. 5,No.3, May 2012.

5. Ritika Mishra and Navjot kaur, "Design and Implementation of Online Punjabi Spell Checker Based on Dynamic Programming" www.ijarcsse.com/docs/papers/Volume_3/8.../V3I8-0308.pd

6. G.S. Lehal and M. Bhagat, " Error Pattern in Punjabi Typed Text," *Proceedings of International Symposium on Machine Translation, NLP and TSS*, 2004, 128-141.

7. Huizhong Duan and Bo-June (Paul) Hsu, " **Online Spelling Correction for Query Completion".**

8**.** Amit Sharma &Pulkit Jain, "*Hindi Spell Checker*", Indian Institute of Technology Kanpur, April 17, 2013

9. li Zhao Dept of Compr Sci& Engr, Xi'an Technological University , "Based on the Phonetic Spelling Correction System Research and Implementation"

10. Dr. R.K Sharma ,"The Bilingual Punjabi English spell checker ," Resource centre for Indain language Technology Solution ,TDIL newsletter.

11. Meenu Bhagat, (2007), "Spelling Error Pattern Analysis of Punjabi
Typed Text", Thesis report, Thapar University, Patiala.

12. "An Analysis of Difficulties in Punjabi Language Automation due to Non-standardization of Fonts", Dharam Veer Sharma Department of Computer Science, Punjabi University, Patiala