Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Review On Error Detection and Error Correction Techniques in NLP

Baljeet Kaur
Department of Computer Science and Engineering,
BFCET Bathinda, India

Abstract- A spell checker is a software tool that detects and corrects any spelling mistakes in a written text. Spell checker analyzes the written text in order to identify misspellings by comparing these words with accepted spellings in database. Most of work has been completed in English and Punjabi language. This paper describes hybrid approaches used for spell checking and correcting system for Hindi language.

Keywords: Errors Classification Error Detection, Error Correction, Dictionary-lookup ,N- Gram Based, Edit-Distance Technique

I. INTRODUCTION

Spell-checking is the process of detecting and sometimes providing suggestions for misspelled words a written text. Ralph Gorin built the first spell-checker for the DEC PDP-10 mainframe computer at Stanford University [1]. 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 misspelled word, and an error corrector that chooses the best correction out of the list of candidate spellings. All spelling checker tools uses a dictionary as a database. Every word from the written text is looked up in the dictionary. When a word is not found in the dictionary, it is detected as an error. In order to correct the error, a spell checker searches the dictionary for words that resemble the erroneous word most. These words are then suggested to the user who chooses the best word that was expected.
There are two main steps in a spell checker. These are Error detection and Error correction.

II. Error Classification

Various studies were performed to analyze the types and the classification of spelling errors.
A real-word error those error words that are acceptable words in the lexicon.

मेरा घर उस और है

For

मेरा घर उस ओर है

और is a acceptable word in Hindi language but it occurs as a error for ओर word.[2]
Non-word error are those error words that cannot be found in the lexicon [3].

ग्यान for ज्ञान

According to Damerau [4] spelling errors are generally divided into two types Typographic errors and Cognitive errors.

A. Typographic errors (keyboard -based)

- These errors are occurring when the correct spelling of the word is known but the word is mistyped by mistake. These errors are mostly related to the keyboard.
- E.g. typing कालघाट for कालाघाट

B. Cognitive Errors

- These are errors occurring when the correct spellings of the word are not known. These errors are occurring when the correct spelling of the word is known but the word is mistyped by mistake due to same pronunciation, called phonetic errors [5]

1. Phonetic Errors: due to same pronunciation but correct spellings are not known.

E.g. हिन्दी and हिंदी.

*2. Homophones Error***:** due to confusion about spelling of two different words.

E.g. आचार(behavior) अचार(pickle).

III. ERROR DETECTION

The main task in spell checking is to detect the errors in written text .There are two techniques for error detection is N-gram analysis and dictionary lookup. The error detection process usually consists of checking to see if an input string is a valid dictionary word or not. Efficient techniques have been devised for detecting such types of errors. Spellcheckers rely mostly on dictionary lookup and *n*-gram techniques.

Dictionary lookup

A dictionary is a list of large number of correct words. Dictionary look up is one of the two principal ways of spelling error detection. Dictionary looks up technique which checks each word of input text for its presence in dictionary. If that word is present in dictionary, then it is a correct word otherwise it is put into the list of error words [6]. Hash tables are the most common used technique to gain fast access to a dictionary. For Input string, one has to compute its hash address and retrieve the word stored at that address in the pre-constructed hash table. If the word stored at the hash address is different from the Input string, a misspelling is flagged. The main disadvantage is the need to devise a clever hash function that avoids collisions. After finding the word incorrect various handcrafted rules are applied to generate the correct spellings of the word by considering the linguistic features of the particular language.

N-gram Analysis [3]

It is a method to detection of incorrectly spelled words in a mass of text. Instead of comparing each entire word in a text to a dictionary, just n-grams are compared with dictionary because comparing each single word with dictionary is a time consuming process. A check is done by using an n-dimensional matrix where real n-gram frequencies are stored. If a non-existent or rare n-gram is detected the word is flagged as a error or misspelled, otherwise not. An n-gram is a set of consecutive characters taken from a string with a length of n. If n is set to one then it is called unigram, if n is two then it is a Bigram, similarly if n is three then the term is trigram. Each string that is involved in the comparison process is split up into sets of adjacent n-grams. The n-grams algorithms have the major advantage that they require no knowledge of the language that it is used with and so it is often called language independent algorithm.

IV. ERROR CORRECTION

when a input word is detected as a error in written text then spelling correction techniques are applied on erroneous word to correct the word or providing correct suggestions for that word .this process is called spell correction.

Rule-based Approach

In this approach handcrafted rules are made by considering the features of the particular language. These rules are applied on the words in the written text which are not found in the database. With the help of these rules the system tries to auto generate the correct spellings of the word which is under observation.

Edit Distance [3]

Edit distance is a simple technique in spell correction. This Simplest method is based on the assumption that the person usually makes few errors if ones, therefore for each dictionary word .the minimal number of the basic editing operations (insertion, deletions, substitutions) necessary to convert a dictionary word in to the non word .the lower, the number ,the higher the probability that the user has made such errors. Through the operation of adding, deleting and modifying, Edit-Distance changes a word into the minimum operating frequency of another word. Edit-Distance set the value as 1 for each change of a word, neglecting the errors arising from the user's wrongly operating habits. For example, the

Edit-Distance is 2 of the wrong word "weke" and the correct word "weak". Suppose we have two strings x,y
e.g. x = kitten
y = sitting
And we want to transform x into y.
We use edit operations: 1. insertions
2. Deletions
3. Substitutions

k i t t e n
s i t t i n g
1st step: kitten →sitten (substitution)
2nd step: sitten→sittin (substitution)
3rd step: sittin→sitting (insertion)
Edit distance is useful for correcting typographic errors, since these are often of the same kind as the allowed edit operations. This algorithm is not good for correcting phonetic spelling errors.

N-gram-Based Techniques

N-gram models can be imagined as placing a small window over a text, in which only *n* words are visible at the same time. The model in which we only look at one word at a time is called unigram model. In similar fashion, a bigram shows two words at a time on a window. N-grams can be thought of n number of words under observation. N-grams can be used in either without a dictionary or together with a dictionary [3]. Used without a dictionary, n-grams are employed to find in which position in the incorrect word the error occurs. If there is a unique way to change the incorrect word so that it contains only valid n-grams, this is taken as the correction. The performance of this method is low. Its main advantage is that it is simple and does not require any dictionary. Together with a dictionary, n-grams are used to define the distance between words, but the words are always checked against the dictionary. This can be done by using several ways, for example check how many n-grams the misspelled word and a dictionary word have common, weighted by the length of the words.

V. EXISTING WORK

In English language most of work has been completed. A common spell checker is available in Microsoft word. In Hindi and Punjabi language small amount of work has been completed. In Hindi language some spell checking applications like **हिंखोज** [7] and **spell guru** [8] are available. Spell guru is paid software.

In Punjabi language, there is very small amount of work is completed in this region. There are two spell checkers for Punjabi language i.e. "Akhar" and "Sudhaar" [3]. **Akhar** is paid software that is not available free for its use to everybody and Sudhaar spell checker is a desktop application. **Raftaar [9]** is a recently developed spell checker for Punjabi language. It is an spell checking online application developed in an ASP.NET language

VI. CONCLUSION

In this paper we have surveyed the area of spell correction and error detection techniques. Existing work related with spell checkers in Hindi and Punjabi language is also discussed. In future I will implement a Hindi spell-checker by using dictionary lookup and edit-distance based technique with more accuracy.

REFERENCES

[1] Michael Rosner, Department of comp. science 'Advanced topics in NLP'.
[2] Amit Sharma &Pulkit Jain, "Hindi Spell Checker", Indian Institute of Technology Kanpur, April 17, 2013.
[3] Neha Gupta, Pratistha Mathur, Spell Checking Techniques in NLP: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X Volume 2, Issue 12, December 2012,
[4] "F.J Damerau (1964)"A technique for computer detection and correction of spelling error", Communicaation ACM.
[5] Li Zhao, 'Based on the Phonetic Spelling Correction System Research and Implementation'.
[6] 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
[7] http://dict.hinkhoj.com/spell-checker/
[8] http://www.bhashagiri.com/
[9] Ritika Mishra, Navjot Kaur, Design and Implementation of Online Punjabi Spell CheckeBased on Dynamic Programming, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 8, August 2013

