International Journal of Computer Trends and Technology- volume4Issue3- 2013

A Survey of Spelling Error Detection and Correction Techniques

Ritika Mishra[1], Navjot Kaur[2]

[1] *Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.*

[2] *Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.*

Abstract— Spelling Correction is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. Spell Checker is an application program that flags words in a document that may not be spelled correctly. Spell Checker may be stand-alone capable of operating on a block a text such as word processor, electronic dictionary. When some text is given as an input to spell checker, it list outs the incorrect words separately by checking their availability in the dictionary. Finally it provides the suggestions for the incorrect words from the dictionary. This survey paper covers almost all the spelling correction techniques.

Keywords— Spell Checker, NLP, Error detection techniques, Error correction techniques.

I. INTRODUCTION

NLP is a form of human-to-computer interaction where the elements of human language, be it spoken or written, are formalized so that a computer can perform value-adding tasks based on that interaction. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyse, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person.

NLP has many applications; they include Automatic Summarization, Machine Translation, Parsing, Information Retrieval, Optical Recognition, and Question Answering.

II. SPELLING CORRECTION

It is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. In computing, Spell Checker is an application program that flags words in a document that may not be spelled correctly. Spell Checker may be stand- alone capable of operating on a block a text such as word-processor, electronic dictionary. Spelling errors can be divided into two categories: Real-word errors and Non-word errors. Real-word errors are those error words that are acceptable words in the dictionary. Non-word errors are those error words that cannot be found in the dictionary.

III. ERROR DETECTION TECHNIQUES

A. Dictionary Lookup Technique

In this, Dictionary lookup technique is used which checks every word of input text for its presence in dictionary. If that word present in the dictionary, then it is a correct word. Otherwise it is put into the list of error words. The most common technique for gaining fast access to a dictionary is the use of a Hash Table. To look up an input string, one simply computes its hash addresses and retrieves 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 or is null, a misspelling is indicated.

B. N-gram Analysis Technique

In this, N-grams are n-letter sub sequences of words or strings where n usually is one, two or three. One letter n-grams are referred to as unigrams or monograms; two letter n-grams are referred to as bi-grams and three letter n-grams as trigrams. In general, n-gram detection technique work by examining each n-gram is an input string and looking it up in a precompiled table of n-gram statistics to as certain either its existence or its frequency of words or strings that are found to contain non-existence or highly infrequent n-grams are identified as either misspellings.

IV. ERROR CORRECTION TECHNIQUES

A. Minimum Edit Distance Technique

The minimum edit distance is the minimum number of operations (insertions, deletions and substitutions) required to transform one text string into another. In its

original form, minimum edit distance algorithms require m comparisons between misspelled string and the dictionary of m words. After comparison, the words with minimum edit distance are chosen as correct alternatives. Minimum edit distance has different algorithms are Levenshtein algorithm, Hamming, Longest Common Subsequence.

1) *The Levenshtein algorithm*: This algorithm is a weighting approach to appoint a cost of 1 to every edit operations (Insertion, deletion and substitution). For instance, the Levenshtein edit distance between "dog" and "cat" is 3 (substituting d by c, o by a, g by t).
2) *The Hamming algorithm*: This algorithm is measure the distance between two strings of equal length. For instance, the hamming distance between "sing" and "song" is 1 (changing i to o).
3) *The Longest Common Subsequence algorithm*: This algorithm is a popular technique to find out the difference between two words. The longest common subsequence of two strings is the mutual subsequence.

For instance, if i= 6750ABT4K9 and j= 0069TYA5L9 then LCS= 650AT9.

B. Similarity key technique

In this, Similarity key technique is to map every string into a key such that similarly spelled strings will have similar keys. Thus when key is computed for a misspelled string it will provide a pointer to all similarly spelled words in the lexicon. A very early, often cited similarity key technique, the SOUNDEX system.

1) *Soundex Algorithm*: This algorithm is used for indexing words based on their phonetic sound. Words with similar pronunciation but different meaning are coded similarly so that they can be matched regardless of trivial differences in their spelling.
2) *The SPEEDCOP System:* It is a way of automatically correcting spelling errors-predominantly typing errors in a very large database of scientific abstracts. A key was computed for each word in the dictionary. This consisted of the first letter, followed by the consonants letters of the word, in the order of their occurrence in the word, followed by the vowel letters, also in the order of their occurrence, with each letter recorded only once.

The Soundex code and SPEEDCOP key are ways of reducing to a manageable size the portion of the dictionary that has to be considered.

C. Rule Based Technique

Rule Based Techniques are algorithms that attempt to represent knowledge of common spelling errors patterns in the form of rules for transforming misspellings into valid words. The candidate generation process consists of applying all applicable rules to a misspelled string and retaining every valid dictionary word those results.

D. Probabilistic Techniques

In this, two types of Probabilistic technique have been exploited.

1) *Transition Probabilities*: They represent that a given letter will be followed by another given letter. These are dependent. They can be estimated by collecting n-gram frequency statistic on a large corpus of text from the discourse.
2) *Confusion Probabilities*: They are estimates of how often a given letter is mistaken or substituted for another given letter. Confusion probabilities are source dependent because different OCR devices use different techniques and features to recognize characters, each device will have a unique confusion probability distribution.

E. N-gram Based Techniques:

Letter n-grams, including tri-grams, bi-grams and uni-grams have been used in a variety of ways in text recognition and spelling correction techniques. They have been used by OCR correctors to capture the lexical syntax of a dictionary and to suggest legal corrections.

F. Neural Net Techniques:

Neural nets are likely candidates for spelling correctors because of their inherent ability to do associative recall based on incomplete or noisy input.

1) *Back Propagation Algorithm*: This algorithm is the most widely used algorithm for training a neural net. A typical back propagation net consists of three layers of node: input layer, an intermediate layer, an output layer. Each node in the input layer is connected by a weighted link to every node in the hidden layer. Similarly each node in the hidden layer is denoted by a weighted link to every node in the output layer. Input and output information is represented by on-off patterns of activity on the input and output nodes of the net. A 1 indicates that a node is turned on and 0 indicates that a node is turned off.

V. CONCLUSION

This paper presented a study on various Spelling Correction techniques. We have discussed various types of spelling errors. Considerable work has been done in the area of English and a related language but there is very little work done in Hindi language. So, In future, we will design and implement new Spell Checker for Hindi language.

REFERENCES

[1] 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.

[2] Youssef Bassil & Mohammad Alwani, "*Parallel Spell-Checking Algorithm Based on Yahoo! N-Grams Dataset*," International Journal of Research and Reviews in Computer Science, Vol. 3, No. 1, February 2012.

[3] Jesus Vilares & Manuel Vilares, "*Textual Spelling Correction: Managing Misspelled Queries in IR Application*," Issue 8, October 2010.

[4] Rupinderdeep Kaur & Parteek Bhatia, "*Spell Checker for Gurmukhi Script*," Thapar University, Issue June, 2010.

[5] Meenu Bhagat, "*Spelling Error Pattern Analysis of Punjabi Typed Text*," Thesis Report, Thapar University, Issue 2007.

[6] Rupinderdeep Kaur and Parteek Bhatia, "*Design and Implementation of SUDHAAR-Punjabi Spell Checker*," International Journal of Information and Telecommunication Technology, Vol. 1, Issue 1, 2010.

[7] E.M. Riseman and A.R. Hanson, "*A Contextual Post Processing System for Error Correction using Binary N-grams*," IEEE Transactions on Computer, pp. 480-493.

[8] B.B. Chaudhuri, "*OCR Error Correction of an Inflectional Indian language using Morphological Parsing,*" TDIL Newsletter.

[9] Dr. Sanghamitra Mohanty, "*Analysis and Design of Oriya Morphological Analyser: some Tests with OriNet,*" TDIL Newsletter.

[10] Davidson, Leon, "*Retrieval of Misspelled Names in an Airlines Passengers record System,*" Communications of the A.C.M, pp. 169-171.

[11] Damerau, F.J., "*A technique for computer detection and correction of spelling errors*", Comm. AC 7(3):171-176, 1964.

[12] Hamming, Richard W., "Error detecting and error correcting codes", Bell System Technical Journal 29 (2): 147–160, 1950.

[13] Kukich, K., "*Techniques for automatically correcting words in text*", ACM Computing Surveys, 24(4), 377–439, 1992.

[14] Jurafsky D., Martin J., "*Speech and Language Processing*", Second Edition, Prentice Hall, 2008.

[15] Joseph J. Pollock and Antonio Zamora, "*Automatic spelling correction in scientific and scholarly text*", Commun. ACM, 27(4):358–368, 1984.

[16] Peterson James, "*Computer Programs for Detecting and Correcting Spelling Errors*", Computing Practices, Communications of the ACM, 1980.

[17] Manning, Raghavan, Schütze, "An Introduction to Information Retrieval", Cambridge University Press, 2008.