ERROR DETECTION AND CORRECTION IN URDU TEXT USING FINITE STATE AUTOMATA

[1]Kanwar Abrar [2] Yousaf Saeed, [3] Muhammad Munwar Iqbal
[1]National College of Business Administration & Economics (NCBA&E) Lahore, Punjab, Pakistan
Email: kanwarabrar@gmail.com
[2]National College of Business Administration & Economics (NCBA&E) Lahore, Punjab, Pakistan
Email: yousafsaeed@ncbae.edu.pk
[3] Department of Computer Science, Virtual University of Pakistan
munwariq@gmail.com

ABSTRACT - *Urdu is most widely used language for communication among people in Pakistan. Detection and correction of wrongly spelled Urdu words are usually done by checking the availability of the particular word in lexicon (dictionary) of Urdu language. In case, the particular word is not found in the lexicon, the misspelled word is corrected by suggesting different correct words within the specific edit distance from the wrongly spelled Urdu word using finite state automata (FSA). It is found that automatic detection and correction of Urdu words in Urdu word processor applications are currently not available. Urdu is morphologically very rich language, so the process of detecting misspelled words in Urdu language is relatively difficult as compared to English. Spell checking and correction of Urdu words can be very beneficial for Urdu word processing applications on different platforms in several different ways that include chatting, short messaging services and emails. In this research article, author proposed a framework for Urdu word processor applications that can automatically detect and correct errors in Urdu text by generating suggestions to the user. Automatic detection and correction of wrongly spelled words using finite state automata results in the improved efficiency of word processor applications.*

Keywords: Urdu, Finite State Automata (FSA), Minimum Edit Distance, Morphological Analysis. Natural Language Processing (NLP), Optical Character Recognition (OCR)

I. INTRODUCTION

Urdu is widely used language which is written and spoken by almost hundred million people of South East Asia including 20 countries of the world [7]. Urdu is declared as the national language of the Pakistan and considered one of the state languages in India [9]. A lot of Urdu based computer applications are being developed now a days. So, the process of detection and correction of wrongly spelled words in Urdu becomes highly important now days in large number of applications such as text editors, OCR systems, chatting applications and etc. According to [1], spell checking is a process of comparing an input word with a complete master list of acceptable words and rejects those words from the input that have no match in the master list.

In this proposed work, we use finite state automata techniques to detect wrongly spelled words and generate various corrections for this word. For detecting a wrongly spelled word, a process of matching the input word is performed alphabetically with the root words in the dictionary [8]. This dictionary is stored in the form of finite state automata. Each legal word in the dictionary has a start and final state. For example, in Urdu, a word " وطن " is compared in the dictionary, if this word is correctly matched in lexicon then this word will be considered a correctly spelled word. Similarly, a word " وتن " will be considered a wrongly spelled word because we will not be able find a correct match for this word in Urdu lexicon.

This proposed framework will check the spellings of Urdu words; authenticate these words and if any error is found, provide the list of corrections to the user. In case, a legal word is not available in the dictionary or no correction available for the input word, user will have the choice to add this word in the lexicon.

II. LITERATURE REVIEW

Spell checking in any language is an important and essential feature of any large word processor application. A lot of research has been conducted and still in progress to make this process more efficient and to develop this system for different languages. Various approaches of spell checking for English and other international languages have been identified so far. Spell checking process is usually performed in three basic steps which are detecting the error, correcting the error and ranking [2]. Currently for English language, hashing, binary search trees and FSA based techniques are used for error detection and similarity key, neural network and edit distance based approaches are being used for correction of wrongly spelled words [4]. Currently, in spell checking procedure, an electronic dictionary is maintained for any particular language having all potential words of the language. The input word is checked within the dictionary; if the word is not available in the dictionary then possible corrections are generated [10]. A basic flow of currently implemented spell checking system is given in the Figure 1.

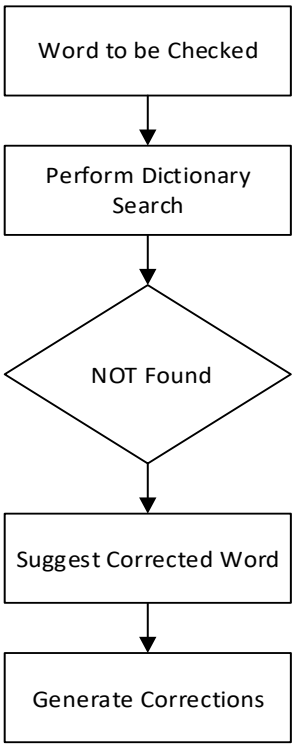

Figure 1: Current Spell Checking System

III. PROPOSED SYSTEM

The proposed system consists of different components based on the input word that the user is typing and based on such input, the proposed framework components activates. Figure 4 indicates the components of the proposed framework and these components are discussed in the following sections.

A. FSA MAPPING

Urdu dictionary is stored in the form of Finite State Automata. Finite state automata are used to represent a language that consist a set of Urdu strings and every string is a combination of symbols from Urdu alphabets. Each legal word in the lexicon has an initial state and a final state [6]. Figure 2 represents the example of mapping of Urdu words in lexicon using finite state automata.

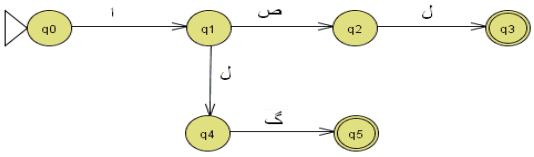

Figure 2: Representation of words "اصل" and "الگ" in Urdu lexicon using FSA

Each input word is mapped into a finite state automaton and this FSA is then compared with the lexicon. Figure 3 represents the mapping of wrongly spelled Urdu word " وتن " using FSA.

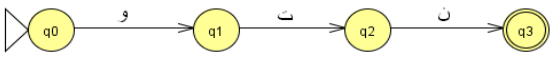

Figure 3: Representation of wrongly spelled Urdu word " وتن " using FSA

B. ERROR DETECTION SCHEME

The simple way to detect wrongly spelled words is to compare the dictionary for the input word. If the input word does not match in the lexicon, morphological analysis is performed to strip the root word from suffix/ affix.

- **Dictionary Checking**

There are several dictionary search approaches available such as hashing, binary search trees and finite state automata. In this work, we use finite state automata based technique to search the lexicon for detecting misspelled words. We construct finite state automata for the input word and compose this input on the Urdu lexicon. This composition provides an intersection of the words that are available both in the input and lexicon [8]. This lexicon is stored in the form of finite state automata. The representation of dictionary using finite state automata makes the process of error detection more efficient.

We take the input FSA and compare it with the dictionary. If we got a match on the arc that is leaving from the starting state, we cross this arc and progress to the next state. If the input word is valid then there will exist a path from starting state to the accepting state in the lexicon. On the other hand, if we will not be able to find a correct match in the lexicon for the particular word, it means that the word is either misspelled or inflected from some root word. For languages such as Urdu which are morphologically very rich, only lexicon checking will not be appropriate for detecting wrongly spelled words. So we need to perform morphological analysis for effective spell checking.

- **Morphological Analysis**

Morphology is the study of basic structure of words. The main component of any word is known as morpheme. According to [5], morpheme can be defined as the smallest component of any specific language that carries the meaning or any information. Every word in any particular language is formed of one or multiple morphemes. For example, in Urdu, some of the words with their corresponding morphemes are shown in Table 1.

Table 1: Morphological Analysis of Urdu Words

Word	Corresponding Morpheme	Addition (Suffix/Affix)
کرسیاں	کرسی	یاں
تمیزدار	تمیز	دار
نالائق	لائق	نا

Morphological analyzer extract the root word by removing the suffixes or affixes attached to the input word [3]. This root word will be checked in dictionary and if we still not able to find a correct match then this word is considered as wrongly spelled word.

C. ERROR CORRECTION SCHEME

In this phase, spell checking system find different words from the lexicon that are some way similar to misspelled words. To correct a wrongly spelled word, first of all we have to find a list of correct words that are close to the

misspelled word. Then, we have to rank these suggested words and select the best candidates those will be suggested to user for correction. There are different correction techniques are available such as minimum edit distance, similarity key technique, probabilistic approach, rule based approach and etc. In this work we used minimum edit distance technique for correcting the misspelled words.

- **Minimum Edit Distance**

Minimum edit distance is the most common and known approach for correcting the wrongly spelled words [4]. In this technique minimum edit operations are performed to change a string in other string. Almost 80% of the spelling errors occur because of single letter error [1]. These errors can be classified as insertion, deletion, substitution and transposition.

Table 2: Urdu Word Errors based on Damerau's work

Type of error	Original Word	Words with Spelling error
Insertion	اصل	اصصل
Deletion	جادو	جاد
Substitution	جیسا	قیسا
Transposition	بلند	بنلد

A matrix of m x n dimensions is taken in this approach; here m represents the length of string 1 and n represents the length of string 2. String 1 is mapped on the first row of the matrix and string 2 is mapped on the first column of the matrix [8]. By using minimum edit distance approach difference of edit operations are placed in every other matrix's cell. In this approach, the wrongly spelled Urdu word is checked with all the words in lexicon to find whether the spelling error could be removed by applying any of the operation (insertion, deletion, substitution and transposition) and edit distance between two strings is measured. For example, the edit distance between "جنگ" and "حنگ" is 1. Here "ج" is substituted for "ح" to correct the word.

- **Correction Suggestions**

In this phase, a list of correct words available in the lexicon within a specified edit distance from the wrongly spelled word is suggested to user. User can select the appropriate word from the list to correct the word.

D. DICTIONARY UPDATION

If a particular legal word is not available in the lexicon, this proposed framework provides a mechanism to add this word in the dictionary. Similarly, user can add any non-word in the lexicon according to his/her requirement to make the lexicon more customized.

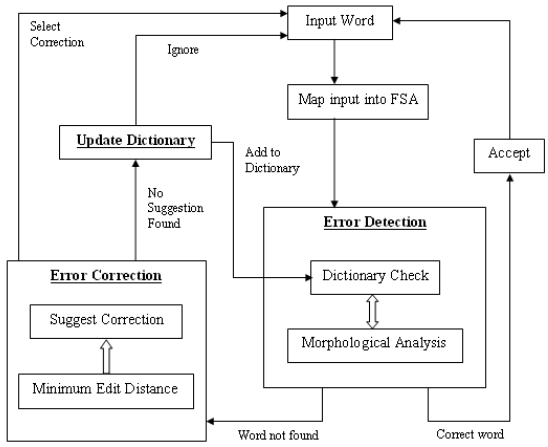

Figure 4: Proposed Framework for Error Detection and Correction of Urdu Text

IV. CONCLUSION AND FUTURE WORK

The proposed framework provides a finite state automata oriented mechanism for detection and correction of wrongly spelled words. Using this mechanism, a wrongly spelled Urdu word can be detected by comparing it with the lexicon. After detecting a wrongly spelled word, different corrections are suggested to user using minimum edit distance technique. If a correct word is not available in the dictionary, user can add the word in the dictionary to make the lexicon richer. Similarly, user can add a non-word in the dictionary if he/she thinks that the particular word should be the part of lexicon. There is still possibility to optimize the process of error detection and correction of Urdu words. Spell checking is the basic process for Urdu grammar checking and part of speech systems. On the basis of the proposed framework, development of a system for Urdu grammar checking can be considered in future.

REFERENCES

[1] F. J. Damerau, "A technique for computer detection and correction of spelling errors" Communications of the ACM, 7, 3, 171-176, 1964.

[2] T. Naseem, S. Hussain, "A Novel Approach for Ranking Spelling Mistakes in Urdu", Language Resources and Evaluation, 2007.

[3] S. Hussain, "Finite-State Morphological Analyzer for Urdu", NUCES, MS Thesis, 2004.

[4] H. L. Liang, "Spell Checker and Correctors: A unified treatment", University of Pretoria South Africa, MS Thesis, 2008.

[5] W. O'Grady, V. P. Guzman, "Morphology: the analysis of word structure" in Contemporary Linguistics: An Introduction, 1997.

[6] A.S. Pillai, "Spell Checker for Tamil using Finite State Automata", Hindustan University, 2010.

[7] N. Durrani, S. Hussain, "Urdu word segmentation ", In the Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Pages 528-536, 2010.

[8] A. Hassan, S. Noeman and H. Hassan, "Language Independent Text Correction using Finite State Automata" In the Proceedings International Joint Conference on Natural Language Processing (IJCNLP), 2008.

[9] S. Hussain, "Resources for Urdu Language Processing", Proceedings of the 6th Workshop on Asian Language Resources, 2008.

[10] K. U. Schulz, S. Mihov, "Fast string correction with Levenshtein automata", International Journal on Document Analysis and Recognition, Volume 5, Issue 1, pp 67-85, 2002.