Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 11, December 2017 – www.caeaccess.org

Spelling Error Assessment Module for Answer Evaluation Process

Amit S. Walters
PhD Research Fellow
Department of CS & IT
SHIATS, Allahabad

Rajendra K. Isaac
Professor, Faculty of
Engineering & Technology
SHIATS, Allahabad

Ajit Paul, PhD
Professor, Head of the
Department of Mathematics and
Statistics, SHIATS

ABSTRACT
This paper discusses methods and algorithm to find various types of spelling errors occurs during writing a word. The examples included in this paper are based on the analysis of writing samples collected for the answer evaluation and other researchers discussed on their papers. Algorithm used in the model works on the different methods for matching the string with list of known word and reassemble word as string with gapes between letter when there is no exact match for best matched word. The module assesses the spelling error through the process of finding gaps in the string of letters.

Keywords
Spelling errors, Causes of spelling errors, Type of spelling error.

1. INTRODUCTION
The knowledge of the subject, writing ability or ability to express the topic and the grammatical mistakes are considered as important factors in the evaluation process of answer script. But in addition to these factors spelling errors gives a vital impact on the final awarded marks. Generally spelling errors in writing a word are classified into two groups (Kuckich, 1992) — typographic and cognitive. Cognitive type of error occurs when the person who is writing the sentence is not familiar with the spelling of the word but can pronounce it and tries to write the word according to the phonetic sound. In these type of error the wrongly spelled word have the similar pronunciation (phonetic sound) as the correct word (for example writing advise as adwese). Typographic or writing errors are mainly occurs due to the mistake in writing or typing through the keyboard; for example substitution or change of letters because on the keyboard keys for these are very close. Damerau (1964) said that 80% of misspelled words that are non-word errors and are the result of insertion of one letter or transposition of one letters.

It can be observed with the help of proportion chart Figure.1 given by cook (1999) that 59% of mistakes are due to the omissions or additions of letter, also it can be observed that omissions or additions of vowels has higher proportion 37% than of consonants 22% respectively. Secondly substitution of letter with only 30%, in substitution of vowels 18% and consonants 12% there is no big difference. Change in position (reversal) 5% and sound based spelling mistakes 6% are much lower [1].

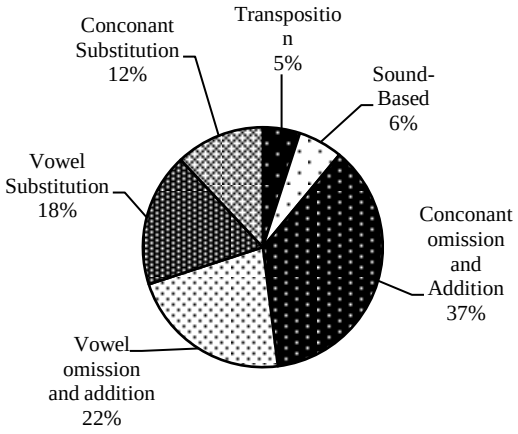

Figure 1: Proportions of spelling mistakes (Cook: 1999)

Further, relationship between pronunciation of word and writing correct spelling Sailaja (2009) observes that a number of English words (often mispronounced) are merely heard, thus it is common to see misspellings across the country. Module is designed to assess the spelling error through the process of finding gaps in the string of letters.

2. LITERATURE REVIEW
On reviewing the past research literature it was observed that it is difficult to put all spelling errors into neatly defined categories and may not be considered as same error. Ott (2007) has discussed that misspelling analysis is time-consuming and often reliant on judgment and it is not exact science – for different types of errors different explanations can be given. The error can be categorized effectively if we know the cause of errors. Krishnamurthy (1978) and Sailaja (2009) have talked about the errors of spelling pronunciation of English word in Indian, which in turn influences the spellings of the learners[2]. Many words like “lovely”, “dedicated”, “irregular”, “nicely” , etc. are very common to hear people, sometimes, don’t ever look up the spellings of the words they hear in their environment. Other common types of spelling errors are the large number of homophonous words in the English language with entirely different spelling[5]. There are many other known and unknown reasons for the occurrence of spelling errors in English. We discuss some them in this paper with the category of spelling errors.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 11, December 2017 – www.caeaccess.org

3. MATERIAL AND METHODS

This section describes different algorithms, methods and process used in this model.

3.1 Problem Identification

It was observed that error in the spelling plays an important role in the evaluation process of answer script. It was discussed that there are different algorithm and methods to find the suggested word for the misspelled word, further it was discussed that all the error couldn't be considered as the same. The severity of error was required to be judged in the evaluation process of answer script, this decision making process requires exact image of the error.

3.2 Methodology used:

The proposed module (Figure 2) evaluates the given word mainly by two methods to provide the spelling error information, which helps in finding the severity of the spelling error.

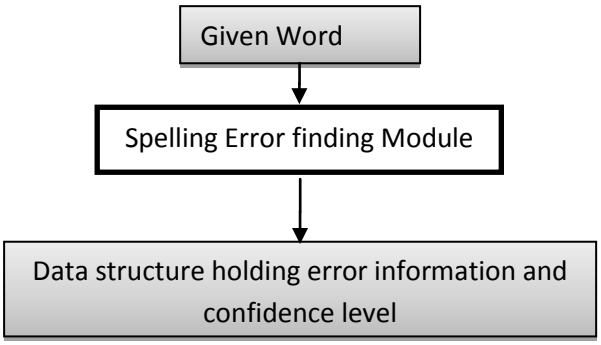

Figure 2: Flow of the module

3.2.1 Method 1

The algorithm used in this method is based on two-dimension array of letters to find mismatch and gapes between letters. In this method it is assumed that the given word is currently spelled and module tries to find match from the word list with +/- one-letter words.

		Q							
		M	**A**	**N**	**M**	**O**	**H**	**A**	**N**
	M	**1**	**0**	**0**	**2**	**0**	**0**	**0**	**0**
	A	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**
	N	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
P	**M**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
	O	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**
	H	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
	A	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**
	N	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**

Figure 3: Correctly Spelled Word in two-dimension array.

To find the correct match of the letters with the highest possible accuracy, a two-dimensional array is allocated. With the row r and column c entry of each letter is denoted here by Erc. There is one column for each character in sequence P, and one row for each character in sequence Q (Figure 3). Thus, if our word Q is of sizes n and word P is of size m, the size of an array used is M(nm).

The algorithm progresses for the computation of error in the given two words P and Q. Each latter P(r) = 1 … n and Q(c) = 1 … m. are matched. The first case if P(r) = Q(c) the value for the corresponding entry Erc is changed to 1. The second case were P(r) ≠ Q(c) the value for the corresponding entry Erc is changed to -1. The third case (Figure 4.1) were P(r) ≠ Q(c) but P(r) = Q(c+1) one blank row is added at the place of P(r) and value for the corresponding entry Erc is changed to -1 (Figure 4.2).

		Q							
		M	**A**	**N**	**M**	**O**	**H**	**A**	**N**
	M	**1**	**0**	**0**	**2**	**0**	**0**	**0**	**0**
	E	**0**	**-1**	**0**	**0**	**0**	**0**	**1**	**0**
P	**N**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
	M	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
	H	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
	A	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**
	N	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**

Figure 4.1: The third case P(5) = Q(5+1)

		Q							
		M	**A**	**N**	**M**	**O**	**H**	**A**	**N**
	M	**1**	**0**	**0**	**2**	**0**	**0**	**0**	**0**
	E	**0**	**-1**	**0**	**0**	**0**	**0**	**1**	**0**
P	**N**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
	M	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**
	-	**0**	**0**	**0**	**0**	**-1**	**0**	**0**	**0**
	H	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**
	A	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**
	N	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**

Figure 4: One blank row is added at the place of P(5).

3.2.2 Method 2

The algorithm used in this method is based on three separate arrays of letters first for given word **P**, second for the word from dictionary database **Q** and third for the new assembled word **N**. In this method it is assumed that every given word is wrongly spelled scrambled word [4] and finds all possible latter combination of the listed word Q. It finds combination with +/- 1 latter, which can match with a word in the dictionary database. Misspelled word 'MENMHAN' has 7 letters. This method tries to find all possible combination of 6 – 8 latter words in the database.

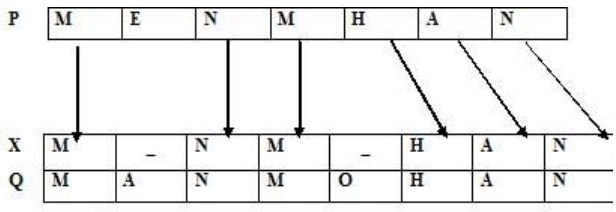

Figure 5: Word Assembly in new array.

The algorithm progresses for the computation of error in the given two words **P** and **Q**. Each latter P(p) = 1 … n and Q(q) = 1 … m. are matched and stored in the third array X(x). The first case were P(p) = Q(q) the letter corresponding Q(q) is placed in the array X(x). The second case were P(p) ≠ Q(q) the value for the corresponding entry X(x) is changed to symbol '-'. The third case were P(p) ≠ Q(q) but P(p) = Q(q+1) the symbol '-' is placed in the array X(x) and with increase in q = q+1 and x= x+1 the letter corresponding

Q(q) is placed in the array X(x). Figure. 5.

In both the method final resultant word is constructed padded with the symbol '-' to represent the location of the error, the confidence level and error code, which shows how close the words are calculated depending on the number of errors, location of error, type of error and word length.

3.2.3 Spelling Error Information

A data structure designed to hold the information of spelling error.

3.2.3.1 Confidence Level

The Confidence level shows how close the given word is with the word tested from the list of words. It is calculated depending on the number of errors, location of error, type of error and word length.

3.2.3.2 Error Code

The error code is an array of bytes. Each byte holds the count of error and type of error on different locations. It is further used as a number, 0 in case there is no error.

3.2.3.3 Padded Word

The padded word is a string which holds the new word formed during the process with '-' symbol in place of mismatch.

4. ANALYSIS AND DISCUSSION

Following are the findings during the experiment.

- **Is the given word is in the dictionary database**: It was observed during the test that if word was found in the dictionary database it returns the pointer to the structure with confidence level 100% on exact match Table 1.
- **Number of letters omitted or added wrongly:** It was observed during the test that if the module do not finds the word in the dictionary database it finds nearly matched word and returns the pointer to the structure with confidence level less than 100% depending on the number of mismatched letters, error code grater then 0 and word padded with '-' on the location of the error (Table 1).
- **Location of wrongly added or omitted letters:** It was observed during the test that if the module do not finds the word in the dictionary database it finds nearly matched word and returns the pointer to the structure with confidence level less than 100% depending on the location of the mismatched letter. Error code grater then 0 (Table 1).
- **Type of letters omitted or added wrongly:** It was observed during the test that if the modules do not finds the word in the dictionary database it finds nearly matched word and returns the pointer to the structure with confidence level less than 100% depending on the type of the mismatched letter (vowels / consonant). Error code grater then 0 (Table 1).

Table 1: Value returned by Error finding Module

Given Words	Confidence level	Error level	No. of wrong letter(s)	Best matched word	Padded word
manmohan	100	0	0	manmohan	manmohan
menmohan	87	0.12	1	manmohan	m-nmohan
menmhan	71	0.25	2	manmohan	m-nm-han
manmoha	85	0.75	1	manmohan	manmoha-
nanmohan	85	0.75	1	manmohan	-anmohan
representative	100	0	0	representative	representative
reprasentative	92	0.07	1	representative	repr-sentative
reprasentative	92	0.07	1	representative	repr-sentative
representativ	92	0.43	1	representative	representativ-
pepresentativ	92	0.43	1	representative	-epresentative
bread	100	0	0	bread	bread
bred	75	0.2	1	bread	bre-d
brea	75	1.2	1	bread	brea-
dread	80	1.2	1	bread	-read

On analysing the data shown in the (Table 1) it was observed that confidence level and error level of the spelling error is related with the number of errors, location of error, type of error and word length. It was further observed that word size is directly related to error level. This shows that smaller words, which are easy to pronounce and memories error level is high then the longer word, which is relatively default to memories.

5. CONCLUSION

Spell a word involves memorising the proper phonetic pronunciation of the word. All spelling mistakes cannot be considered same. Level of the mistake depends on the size and complexity of the phonetic pronunciation of the word. Values for analysis in Table 1 shows, that all mistakes may not be taken equal, it can be observed that there is only one error in different sized word the confidence level is different. Confidence level depends on the number of error. The module for the assessment of the spelling mistake works in the same format. Further more methods can be added for batter confidence level.

6. REFERENCES

[1] Cook, V. (1999) – Teaching Spelling http://homepage.ntlworld.com/vivian.c/Writings/Papers/TeachingSpelling.htmt

[2] Sailaja, P. (2009). Dialects of English: Indian English. Edinburgh: Edinburgh University Press.

[3] Ritika Mishra , Navjot Kaur (2013) A Survey of Spelling Error Detection and Correction Techniques International Journal of Computer Trends and Technology volume 4

Issue 3 - 2013 ISSN: 2231-2803 Page 372

[4] F. J. Damerau. (1964) A technique for computer detection and correction of spelling errors. In *Communications of the ACM*, volume 7(3), pages 171–176.

[5] Bebout, L. (1985) 'An error analysis of misspellings made by learners of English as a first and as a second language'.Journal of Psycholinguistic Research 14/6: 569-593.