Automatic Arabic Spelling Errors Detection and Correction Based on Confusion Matrix- Noisy Channel Hybrid System

[1]Hatem M Noaman, [1]Shahenda S. Sarhan, and [2]M. A. A. Rashwan
[1]Computer Science Department, Mansoura University, Egypt
[2] Electronics and Communications Department, Cairo University, Egypt
hatemnoaman@yahoo.com, shahenda_sarhan@yahoo.com, mrashwan@RDI-eg.com

Abstract

Arabic spelling errors occur in different types of documents, such as handwritten by non experienced users, optical character recognition (OCR) documents and machine translated documents. Many researchers had tried to solve this dilemma but till now there is no a radical solution.

This paper proposes a hybrid system based on the confusion matrix and the noisy channel spelling correction model to detect and correct automatically Arabic spelling errors. The proposed system is based on building a robust error confusion matrix using 163,452 pairs of spelling errors, and its corrected form extracted from Qatar Arabic Language Bank (QALP) and using this matrix with language model to generate list of candidates and choose the most appropriate candidate for given misspelled word. Comparing the proposed system results shows that system result outperform other systems results.

Keywords: *Spelling detection and correction, Noisy channel model, Arabic Natural Language Processing*

1. Introduction

The problem of spelling errors is one of the common problems in written text. Text can be generated from different sources either by human as document typing and emailing software, or by machine like optical character recognition (OCR) and machine translation (MT). This increases the need to build robust and effective approaches to detect and correct automatically spelling errors in Arabic text.

Spelling correction system involves two main modules the, first detects the spelling mistake in the written text, while the second corrects the spelling errors. Through the First part, errors are simply detected using a lexicon of correct words, and if any given word in the text is out of lexicon, it will be considered as spelling error, Then errors correction module would generate a list of ordered candidates that could be considered corrections for the misspelled word, while in automatic spelling correction systems only one word is chosen as the correct word.

Many approaches such as substitution rules, n-gram, Noisy Channel Model, distance ranking and more are investigated to handle spelling errors detection and correction problem. In this paper the researchers concentrated on using the noisy channel model which is one of the most widely used approaches. Which treated the misspelled words as if the correct word is distorted during the passing of the communication channel, and our goal is to build a model for this channel and pass every misspelled through this model and try to correct it with special version of Bayesian inference rule. Simply by finding the correct word that can generate this misspelled word (typo) as in equation (1) (Kernighan et al,1990).

$$\dot{w} = argmax_{word}\, P(word|typo) = argmax_{word} P(typo|word) * P(word) \qquad 1$$

correct system proposed by (Kernighan et al,1990) reported accuracy about 87% of only 392 test cases, to ensure robustness of the proposed system for Arabic language which is highly inflective language where each word can have many morphological forms this paper uses two test consists of 841 test cases and 2027 test cases respectively.
System proposed by (Kernighan et al, 1990) estimates word probability P(word) as:

$$P(word) = (freq(word) + .5)/N \qquad 2$$

Where freq (word) is the number of times that the word appears in corpus and N is size of corpus. Proposed system in this paper instead of computing P(word) using freq(word) and adding 0.5 to handle non seen words which will affect the probability of seen words, proposed system computes P (word) using the language model probability of a given word using a corpus of a target language, language model was built using SRILM and uses Modified Kneser Ney smoothing algorithm to estimates probability for unseen words, and P (typo | word) is computed based on a confusion matrix based on the operation performed on the typo to be corrected to the correct word, as in equation (3) (Kernighan et al,1990).

$$P(typo\,|word) = \begin{cases} delete[x,y]/_{count[x\,y]} & if\ delete \\ insert[x\,,y]/_{count[x]} & if\ insert \\ subsitute[x,y]/_{count[y]} & if\ subsitute \\ transpose[x,y]/_{count[x\,y]} & if\ transpose \end{cases} \qquad 3$$

This paper is organized as follows; in section 2 presents an overview of related work in the field of automatic spelling error detection and correction. Section 3 will discuss the training set construction and confusion matrix training procedure. Testing sets and proposed spelling error correction procedure is presented in section 4, proposed system results are presented in Section 5and finally conclusion remarks are stated in section 6.

2. Preliminaries and Related Work

Spelling errors detection and correction in English Language was broadly investigated, and many researches are done with English errors, (Deorowicz, 2005) proposed a system to correct spelling errors based on classifying mistakes and build its substitution rules in order to improve candidate suggestion. While (Schaback, 2007) achieved 90% for first candidate up to 97% for first five candidates based one detecting the errors one various language levels, phonetic level , the character level, word level, syntactic level and semantic level, his system outperforming MS Word, Aspell, Hunspell, FST and Google.

In Arabic, different approaches are applied to solve this problem, (Shaalanet al. , 2003) tried to build a spelling checker tool for Arabic that can capture common errors mistakes for Standard Arabic and Egyptian dialects. (Haddad and Yaseen, 2007) tried to detect and correct non-word in Arabic text using hybrid approaches based on morphological knowledge phonetic bi-gram rules. While, (Shaalan et al., 2010) uses Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) to detect spelling errors and generate candidates using the edit distance algorithm based on the transformation rules.

(Alkanhal et al., 2012)Proposed an approach based on using a lattice search, and an n-gram method to search in a generated list of all possible alternatives for each misspelled word, they generated this list using Levenshtein edit distance.

In addition, (Shaalan et al., 2012) applied Noisy Channel Model with Language models and knowledge-based rules for error correction using 9 million word list. Another work presented by (Attia et al., 2012) uses the same 9 million word list and enhances the language model by analysis the percentage of noise and use the most optimal data set to train their language model, as well analyze the errors types to improve the edit distance ranking algorithm

Commercial tools were developed by Microsoft and used in its products as Microsoft office, which provide correction for common mistakes, but it is very limited (Attia et al, 2012), Also Google tried to improve its search engine query by applying some rules on Arabic spelling mistakes (Hammad, 2010) reported that Arabic search results enhanced by 10% due to spelling error correction. MADAMIRA system (Pasha et al., 2014) can be used to correct some errors like Hamze spelling errors in Arabic, which considered as a common spelling mistake in Arabic typed texts. In this work (Zerrouki et al, 2014) regular expression with word list to detect and correct errors. Additionally, (Nawar and Ragheb, 2014) tried to use probability scored correction rules to maximize F-score of the training input data to handle the problem of spelling correction.(Mostafa et al, 2014) investigated two different approaches first they used a lexicon driven approach using Hunspell as a spell checker and correction tool and the second one is based on SMT systems using Moses with 1 million tokens training corpus, They reported that SMT system was better than Hunspell approach.(Attia et al, 2014) and .(Attia et al, 2015) proposed a hybrid approach that uses CRF for handling punctuation errors and improve the word list and LM parameters, and they also introduce a proposed algorithm of handling merged words errors .

A hybrid approach that combines rule-based linguistic techniques , language modeling and machine translation, as well as an error-tolerant finite-state automata method was proposed by (Bouamor et al , 2015). While (Nouf AlShenaifi et al, 2015) proposed hybrid cascade model uses probalistic models combined with edit-distance approach to solve the problem of Arabic spelling errors correction problem.

The Availability of large number of training examples can be used to a good confusion matrix that can be used within noisy channel model to build robust Arabic Automatic spelling errors correction system and this what the researchers are trying to do in this paper. So in the remaining of this paper we will introduce the main modules of our proposed hybrid system, system results and implementation details.

3. The confusion Matrix- Noisy Channel Spelling Correction hybrid System

3.1. Dataset

3.1.1. Training Data set

Training data set is collected from Qatar Arabic Language Bank (QALP) (Zaghouani et al, 2014) training set. QALP corpus is a set of sentences with different types of errors and their corrections. While there are different forms of errors that are represented in QALP corpus, only edit errors was extracted, which are words with spelling errors and associated with their corrected form, a set of 163,452 pairs of spelling errors and its corrected form was used to train our model. Also extracted 18 common separated prefixes to handle the split common mistakes in Arabic text like writing (عبدالله) as (عبدالله) without space between the two words. Table (1) presents examples for different forms of errors that were used in our training set

Table1: examples for different forms of errors that was used in our training set

Typo	English Translation	Correct word
الى	To	إلى
كذالك	as well	كذلك
الأدميين	human being	الادمين
زوجه	wife	زوجة
لطهيرها	To clean	لتطهيرها
الصينينة	Chinese	الصينية

3.1.2. Testing Data sets

In order to test proposed system accuracy two test sets were used, first one is extracted from Qatar Arabic Language Bank (QALP) (Zaghouani et al, 2014) corpus, this test set is used to find out the system accuracy using words from the same training data source. Second test set consists of 2027 pairs of spelling errors and its corrected form; this set is adapted from (Attia, et al. 2012) which used to compare proposed system results with Attia's reported results.

3.2. The Confusion Matrix Training Algorithm

The First step through the proposed system is to build Arabic spelling errors confusion matrix using the training pairs extracted from QALP corpus, each pair from the training set is presented to the algorithm to select the correction operation that is applied to the typo characters to produce the corrected form of this typo. The confusion matrix training algorithm assumes that there is four operations that can cause any spelling mistake insert, delete,

transpose or replace. After determining the operation, values of the appropriate confusion matrix are modified based on the correction operation as illustrated in Figure (1).

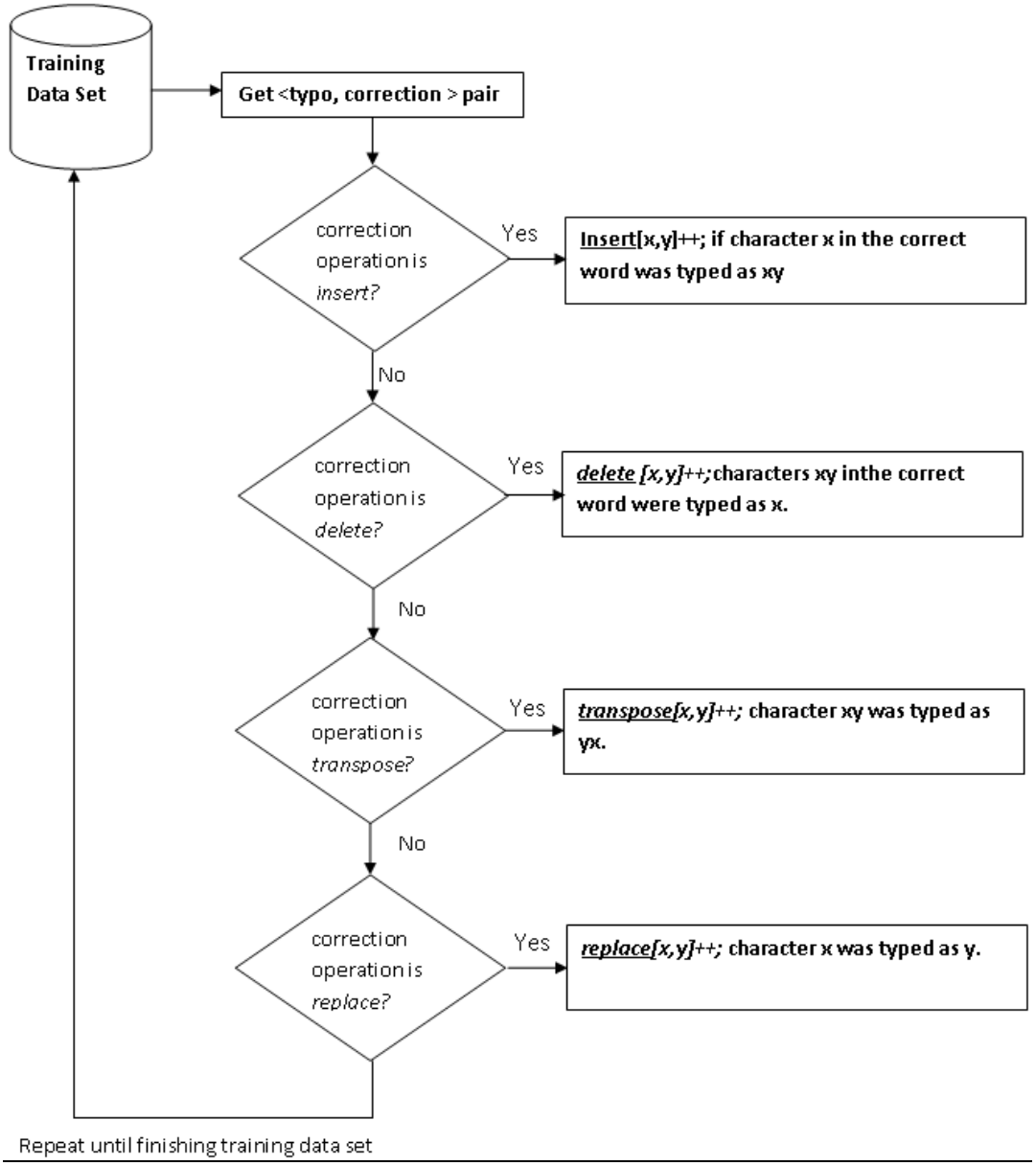

Figure 1: Confusion Matrix Training Algorithm

3.3. The Noisy Channel Spelling Correction Model

If a given word is presented to the system, this word may be correct or have a spelling mistake and in order to decide whether the word contains errors or not, the system checks if the word is existed in a dictionary that contains 355308 entries. If the word appears in the dictionary, it will be in the correct form otherwise the algorithm count this word to have spelling mistakes. Candidates' generation is based on confusion matrixes that are built in the training step. First step is to try to split the word into two parts to check if the word is two words with missing space between them by comparing the first part of the word with a set of

predefined common Arabic prefixes if the word has one of them and the other part is a valid Arabic word, then add this two parts with space between them to the candidates list.

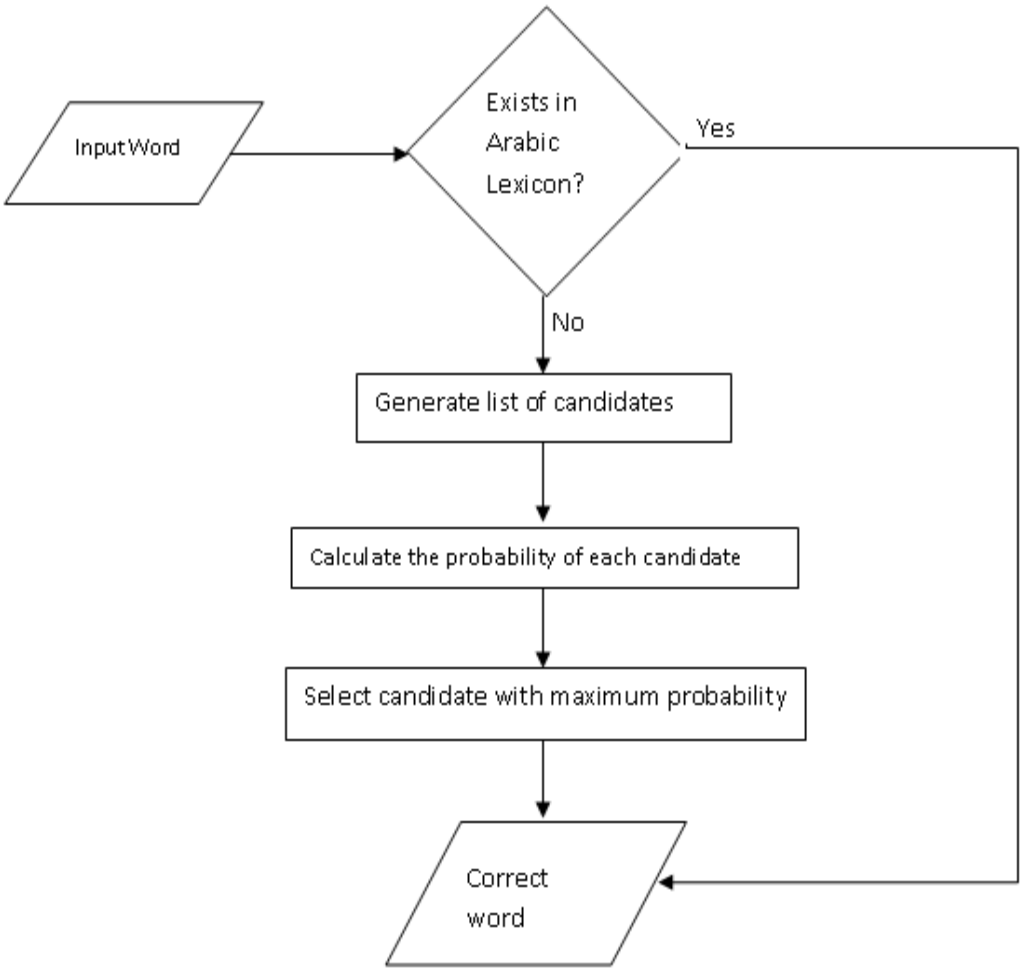

Figure2: Noisy Channel spelling correction model

Next the system tries to get candidates by applying different character-based operations to the word, to generate all valid words after applying delete, transpose, replace and insert operations and calculate the probability if each candidate using the language model and the confusion matrix. If there is no any candidate generated from this step, the system tries to get all candidates with no concern to the confusion matrix, and the candidate probability will be based on the language model only. Figure (2) represents the spelling correction algorithm steps.

As a testing example table 2 introduces a sample "بدروهم" , which contains transpose spelling error, where two characters 'ور' are transposed into 'رو'after introducing this word to the system, the system produces a set of candidates presented in table 2 , all the produced candidates are correct Arabic words, each candidate is associated with its probability, the candidates are ranked based on the candidate's probability and candidate with the maximum probability score will be selected as the correct candidate, as shown in table (2), the selected correct form for "بدروهم" is "بدورهم" that scores 0.02505873.

Table 2: candidates generated by the system for "بدروهم"

Candidate	English Translation	Probability
بدورهم	By their turn	0.02505873
قدروهم	Estimate them	0.001196893
صدروهم	Export them	0.0004580793
بدروم	basement	0.0002626684

4. Implementation

The proposed System was implemented using Microsoft C#.Net ® 2012, while the training dataset was extracted from QALP corpus (Zaghouani et al, 2014), which is resented as plain text files, then the extracted training examples is saved into relational database implemented using Microsoft SQL-Server 2008®, system uses SRILM (v 1.9) to compute the probability of a words from a Arabic corpus collected from online documents composed nearly of 1 million words.

5. Results and Discussion

To test the proposed spelling errors correction algorithm presented in this work, two different test set are used, first one is extracted from Qatar Arabic Language Bank (QALP) (Zaghouani et al, 2014), second test set is adapted from (Attia et al, 2012) work.

Qatar Arabic Language Bank (QALP) test set consists of 841 test cases; each test case contains misspelled word and its correction. Table 3 shows the results obtained using this test set. As noticed from results in table 3 accuracy increased by more than 12% as the generated candidates cutoff limit increased from one candidate to two candidates, the reason for this sharp accuracy improvement is depending on the Qatar Arabic Language Bank (QALP) corpus corrections is context dependent corrections; i.e.: proposed model may give a very reasonable correction but it appears as a second system choice in the generated candidate list, which gives an indication that results may be enhanced if we take the word context factor in the misspelling word correction process .

Table 3: proposed system results using Qatar Arabic Language Bank (QALP) corpus test set

Cutoff limit	Accuracy
1	72.65%
2	84. 90%
3	86.21%
4	86.68%
5	86.92%
6	87.28%
7	87.51%
10	87.87%

In order to compare our results against the work reported by (Attia et al, 2012), the same test dataset was used. This test is helpful and robust because the researchers had reported some results related to using this test dataset and compared their results with three different text authoring software; Google Docs, Open-Office Ayaspell, and Microsoft Word, according to results reported by (Attia et al, 2012), Microsoft Word accuracy was 71.24%, while Open-Office Ayaspell and Google Docs are 41.88% and 17.02% respectively at word type level. The ratio is downgraded at the word token level to 57.15%, 41.86% and 9.32% for the three systems. Attia's system achieved 78.39% with two words cutoff limit and 82.86% when cutoff limit increased to three with the same test dataset.

Table 4 compares between proposed system results and Attia et al work with the same 2027 misspelled words the comparison shows that our system outperforms Attia's results.

Table 4: result comparison between our system and Attia et al 2012 system results

Cutoff limit	Proposed System	Attia et al System
1	85.45 %	-
2	88.75 %	78.39 %
3	89.15 %	82.86 %
4	89.20 %	81.79 %
5	89.39 %	81.03 %
6	89.44 %	80.73 %
7	89.59 %	80.12 %
8	89.64 %	79.58 %
10	89.69 %	78.87 %

Figure 3 illustrates overall system accuracy as cutoff limit is increased using previously stated test sets, figure shows system achieves a noticeable improvement (~12% for Qatar Arabic Language Bank (QALP) test set and ~5% for Attia et al test set) if system suggests two words instead of one suggestion, and the curve after this limit is nearly the same as the cutoff in increased. So it will be better if the user has an option to choose the next generated candidate if he decides that the system selected candidate is not the optimal correct one.

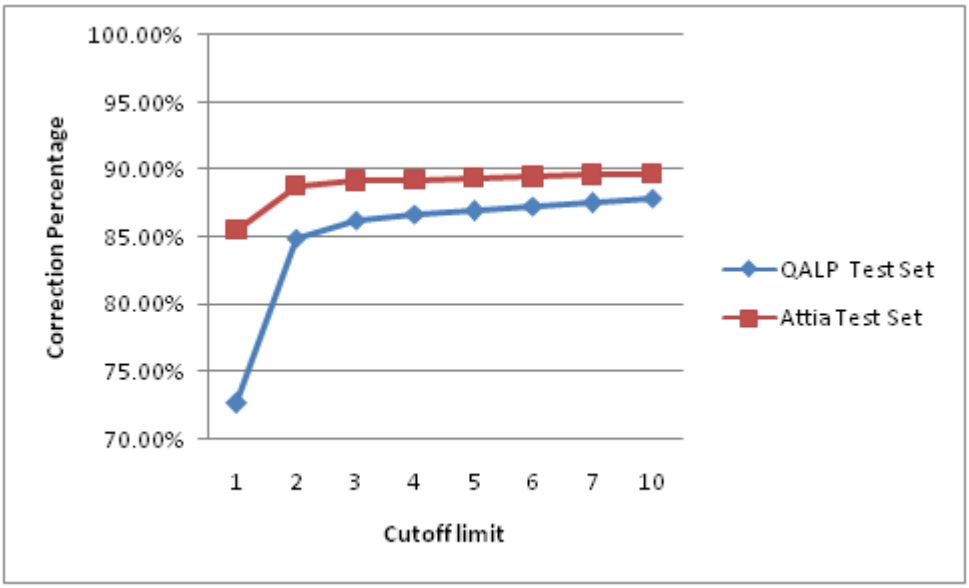

Figure 3: Percentage of words for which a proper correction was found in the top n-generated corrections using Qatar Arabic Language Bank (QALP) and Attia et al, 2012 test set

From results founded by this work we can observe two points:

- Noisy channel model algorithms Hybrid with language model showed notable accuracy when it used to solve Arabic spelling errors correction problem
- The proposed system accuracy had not have any remarkable improvement after two word cutoff limit for the corrected word suggestions

Using the spelling errors correction ratio as performance measure, it's been proved that using Noisy channel model in this process raises the total performance with rates of [7%]. So using Noisy channel model is very useful for correction of Arabic spelling errors in Human and machine generated text documents.

6. Conclusions and Future Work

Automatic correction in Arabic text represent one of the main dilemmas in natural language processing field as Arabic language is one of the most difficult languages to deal with Arabic grammar rules and linguistics.

So in this paper, and confessing the value of the Arabic language the holy Quran language, we presented a hybrid system based on the confusion matrix and Noisy Channel spelling correction model to detect and correct Arabic spelling errors. The overall system accuracy was 85.45 % with the first candidate choice and 89.69 % if we extend the candidates cutoff limit to 10 candidates, which is a good improvement over the state of art approaches.

In the future the researchers are intending to proceed with extending the proposed hybrid model to detect and correct context errors, these types of errors where words are correct in spelling form but wrong within the current context

References

[1] Alkanhal, M. I., Al-Badrashiny, M. A., Alghamdi, M. M., & Al-Qabbany, A. O. (2012). Automatic Stochastic Arabic Spelling Correction With Emphasis on Space Insertions and Deletions. Audio, Speech, and Language Processing, IEEE Transactions on, 20(7), 2111-2122.

[2] Attia, M., Al-Badrashiny, M., &Diab, M. (2014). GWU-HASP: Hybrid Arabic Spelling and Punctuation Corrector1. ANLP 2014, 148.

[3] Attia, M., Pecina, P., Samih, Y., Shaalan, K. F., & van Genabith, J. (2012, December). Improved Spelling Error Detection and Correction for Arabic. In COLING (Posters) (pp. 103-112).

[4] Buckwalter,T.(2002). Buckwalter Arabic Morphological Analyzer. Linguistic Data Consortium. (LDC2002L49).

[5] Deorowicz, S., &Ciura, M. G. (2005). Correcting spelling errors by modelling their causes. International journal of applied mathematics and computer science, 15(2), 275.

[6] Haddad, B., &Yaseen, M. (2007). Detection and correction of non-words in arabic: A hybrid approach. International Journal of Computer Processing of Oriental Languages, 20(04), 237-257.

[7] Kernighan, Mark,D., Kenneth, W., Church, and William, A., Gale. (1990). A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

[8] Mostafa, D., Asbayou, O., & Abbes, R. (2014). TECHLIMED System Description for the Shared Task on Automatic Arabic Error Correction. ANLP 2014, 155.

[9] Nawar, M. N., &Ragheb, M. M. (2014). Fast and Robust Arabic Error Correction System. ANLP 2014, 143.

[10] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., ...& Roth, R. M. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland.

[11] Schaback, J., & Li, F. (2007, January). Multi-level feature extraction for spelling correction. In IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data (pp. 79-86).

[12] Shaalan, K., Aref, R., &Fahmy, A. (2010, March). An approach for analyzing and correcting spelling errors for non-native Arabic learners. In Informatics and Systems (INFOS), 2010 The 7th International Conference on (pp. 1-7). IEEE.

[13] Shaalan, K. F., Attia, M., Pecina, P., Samih, Y., & van Genabith, J. (2012, May). Arabic Word Generation and Modelling for Spell Checking. In LREC (pp. 719-725).

[14] Shaalan, K., Allam, A., &Gomah, A. (2003, October). Towards automatic spell checking for Arabic. In Conference on Language Engineering, ELSE, Cairo, Egypt.

[15] Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh, N., Rozovskaya, A., ..&Oflazer, K. (2014, May). Large scale arabic error annotation: Guidelines and framework. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland.

[16] Zerrouki, T., Alhawaity, K., &Balla, A. (2014). Autocorrection Of Arabic Common Errors For Large Text Corpus. ANLP 2014, 127.

[17] A recent improvement for Arabic searches (http://googleblog.blogspot.com/2010/02/recent-improvement-for-arabic-searches.html) [Accessed : June 28, 2015]

[18] Attia, M., Al-Badrashiny, M., &Diab, M. (2015, July). GWU-HASP-2015@ QALB--2015 Shared Task: Priming Spelling Candidates with Probability1. In ANLP Workshop 2015 (p. 138).

[19] Bouamor, H., Sajjad, H., Durrani, N., &Oflazer, K. (2015, July). QCMUQ@ QALB-2015 Shared Task: Combining Character level MT and Error-tolerant Finite-State Recognition for Arabic Spelling Correction. In ANLP Workshop 2015 (p. 144).

[20] AlShenaifi, N., AlNefie, R., Al-Yahya, M., & Al-Khalifa, H. (2015, July). Arib@ QALB-2015 Shared Task: A Hybrid Cascade Model for Arabic Spelling Error Detection and Correction. In ANLP Workshop 2015 (p. 127).