International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 13, No.1 (Jan-2023)

http://dx.doi.org/10.12785/ijcds/130129

Global Spelling Correction in Context using Language Models: Application to the Arabic Language

Saida Laaroussi[1], Abdellah Yousfi[2], Si Lhoussain Aouragh[3] and Said Ouatik El Alaoui[1]

[1]*ES-Lab, ENSA, Ibn Tofail University, Kenitra, Morocco*
[2]*FSJES Souissi, Mohamed V University in Rabat, Morocco*
[3]*ICES Team, ENSIAS, Mohamed V University in Rabat, Morocco*

Received 6 Oct. 2022, Revised 19 Nov. 2022, Accepted 15 Dec. 2022, Published 31 Jan. 2023

Abstract: Automatic spelling correction is a very important task used in many Natural Language Processing (NLP) applications such as Optical Character Recognition (OCR), Information retrieval, etc. There are many approaches able to detect and correct misspelled words. These approaches can be divided into two main categories: contextual and context-free approaches. In this paper, we propose a new contextual spelling correction method applied to the Arabic language, without loss of generality for other languages. The method is based on both the Viterbi algorithm and a probabilistic model built with a new estimate of n-gram language models combined with the edit distance. The probabilistic model is learned with an Arabic multipurpose corpus. The originality of our work consists in handling up global and simultaneous correction of a set of many erroneous words within sentences. The experiments carried out prove the performance of our proposal, giving encouraging results for the correction of several spelling errors in a given context. The method achieves a correction accuracy of up to 93.6% by evaluating the first given correction suggestion. It is able to take into account strong links between distant words carrying meaning in a given context. The high-level correction accuracy of our method allows for its integration into many applications.

Keywords: Global Contextual Correction, Single Error Correction, Misspelling, Viterbi Algorithm, n-gram Language Model, Edit Distance, Arabic NLP

1. INTRODUCTION

The great expansion of digital linguistic data requires efficient tools for their processing. Language technologies offer resources for producing, transforming, analyzing and researching these data. The relevance of the textual information depends on the processing quality of these tools. The automatic spelling correction is a traditional major component of Natural Language Processing (NLP) [1] applications. For instance, Optical Character Recognition (OCR) [2], machine translation [3], speech recognition [4] and search engines [5] integrate automatic error correction systems. The difficulty of the correction operation often varies depending on the language to be processed. Arabic language presents various challenges in terms of its rich morphology and structure, which makes its analysis more complicated [6]. Arabic is the fifth most spoken language in the world [7]. Many research works have been carried out in spelling correction for Arabic texts. However, these works remains insufficient and they have not reached the level of relevance in comparison with works in other languages such as English.

Automatic correction systems can be classified in two types. Systems that simply offer potential correction suggestions for each detected error, resulting in a manual selection of the desired correction. Other systems can automatically select a single suggestion of correction. The purpose of the autocorrect function is to automatically correct spelling errors in a text without having to return to the user [8]. While spell checkers can help users correct spelling mistakes themselves, they can also help users to choose the correct spelling from a proposed list of (interactive) suggestions [9]. MacArthur [10] et al. and Montgomery [11] found that the spell checker is more effective if the correct spelling is provided in the first three suggestions. But without consideration of a criterion for ranking these correction candidates, users may have difficulty in choosing the correct word among those suggested. It is therefore advisable to keep the list of suggestions as short as possible.

The approaches used for spelling correction can be split into two major categories depending on whether they introduce context or not [12]. The first category is based on a set of rules designed in advance. This type of approach focuses only on the error to be corrected, ignoring its context. It functions as a guide in finding the best suitable

E-mail address:
saida.laaroussi1@uit.ac.ma, yousfi240ma@yahoo.fr, l.aouragh@um5r.ac.ma, s_ouatik@yahoo.com

candidate. The approaches of the second category use, in addition, a model based on the context, because of the insufficiency in general of the a priori error model. This model uses the stochastic methods of Brill and Moore [13]. The introduction of the context can be done via deep neural networks. In practice, spelling correction methods adapt their model to the problem to be addressed by combining two or more models.

This paper investigates a new method, based on stochastic models, for automatic correction of spelling errors applied to the Arabic language. The method takes into account the globality of incorrect word forms seen in their context and examines the semantic links between words and the possible correction candidates, in order to refine the potential global correction. It makes it possible to correct several errors, globally and simultaneously, in a given context, by taking into consideration the respective lists of candidates for correcting all the errors. This method is based on the combination of a priori error model implemented through the edit distance and a context model based on a new estimate of the n-gram language model [14][15] learned from an Arabic corpus. It models the erroneous words as hidden states of a Hidden Markov Model (HMM) and it is driven by the Viterbi algorithm. The main contributions of this work can be summarized as follows:

- We propose an effective spelling correction method using the words context to globally and simultaneously correcting spelling mistakes for Arabic language.

- We exploit a Hidden Markov Model (HMM) to identify the best suggestions of correction of all the misspelled words when they occur in sentences.

- We investigate the Viterbi algorithm using a probabilistic model based on a new estimate of n-gram language models combined with the edit distance.

- We evaluate the performance of our method by conducting a set of experiments on an Arabic corpus and comparing results with other automatic spelling correction systems.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 introduces the global contextual spelling correction. Section 4 describes our proposed method. Section 5 illustrates the experimental results. Finally, Section 6 concludes and outlines the future works.

2. RELATED WORK

The elaboration of automatic spelling correction systems requires understanding the origin of spelling errors [16]. According to Kukich [12], spelling errors can be divided into two types. The first type concerns the cognitive errors, where the user does not know the correct way of writing (new language learner, disabilities...). These errors are user and language specific. Typographic errors are the second type, which is usually related to the keyboard or input device. Fast typing may also cause typographic errors. Approaches dealing with automatic spelling correction can be divided into two major groups, depending on whether they focus only on the error or whether they take into account the surrounding context. Many approaches combine between two models or more depending on the problem to address. We can differentiate between context-free and Context-oriented models.

A. *Context-Free Model*

This model focuses on the error shape, disregarding its context, to generate suggestions of correction [17]. This involves detecting the wrong word and proposing a set of correction candidates that are the most likely similar to the non-word error. The more the best-fit suggestions are proposed first, the more efficient the model is. Researchers, for many decades, have devoted considerable work to enhancing the results referring to this model. In this sense, we cite the following research works.

A study done by Damerau [18] revealed that about 80% of misspelled words occur as a result of a single mistake. He assumes that a word not found in a dictionary has at most one error, which might be a wrong, missing or extra letter or a single transposition. The unidentified input word is again compared to the dictionary, each time testing whether the words match, assuming that one of these errors has occurred. When tested on scrambled text, correct identifications were made for over 95% of these error types.

Kernighan et al. [19] proposed a new program that takes words rejected by the Unix spelling program and provides a sorted list of candidates according to probability scores, based on a noisy channel model. They considered the four edit operations (insertion, deletion, substitution and reversal) and adopted a classic Bayesian argument to find the correction c for the typo t by maximizing Pr(c)*Pr(t/c).

Alwabel [20] proposed a novel error correction mechanism called CoEdit approach to help compilers suggest the most suitable repair for programming errors that occurred as a result of mistyping errors. He employs Four-Way and Editex algorithms [21][22] joined with his approach in order to find repairs to misspelling errors. He finds that using the Editex algorithm with CoEdit is the best choice in the case of finding repairs to programming errors that occur because of spelling errors.

B. *Context-Oriented Model*

The automatic spelling correction based just on the error model remains insufficient and limited. In fact, the context surrounding the erroneous word completes the global meaning of the sentence [23]. Better results could be achieved if additional parameters relative to the context were taken into account. Many research studies have been done in this direction.

A research conducted by Brill and Moore [13] describes an improved error model for noisy channel spelling cor-

rection based on generic string-to-string edits. Unlike the majority of the work conducted in this sense, the research has gone into improving the channel model instead of the source model for spelling correction. The study combines the use of the new improved model and language models to obtain better results. The new error model, without a language model, gives a 52% reduction in spelling correction error rate compared to the weighted Damerau-Levenshtein distance technique of Church and Gale [24]. Whereas, with a language model, the results reach 74% error reduction.

Aouragh et al. [25] have developed a system for correcting spelling errors in the Arabic language based on Levenshtein algorithm and language models. They adapted the Levenshtein distance by adding a weighting based on language models in order to present a solution to filter and refine the correction candidates obtained a priori with the Levenshtein algorithm. The correction rate was order of 95% against the Levenshtein distance that reaches only about 12%.

Dong et al. [26] have conducted research to solve the out-of-vocabulary problem caused by Uyghur spelling errors in Uyghur–Chinese machine translation in order to improve its quality. They assessed three spelling correction methods based on machine translation: BLEU (Bilingual Evaluation Understudy) score, Chinese language model and bilingual language model. They concluded that the best results are achieved in the spelling correction task joined to the machine translation task by using the BLEU score for spelling correction.

Laaroussi et al. proposed new language models based on n-gram models combined with edit distance to deal with spelling errors [14][27]. The advantage of the new models compared to the classic ones consists in taking into account the strong links between distant words in a given context by keeping the model parameters simple. The accuracy obtained, by applying the new models on an Arabic corpus, exceeds 98% for the first ten correction candidates, and 79% for the first correction to suggest.

Nejja and Yousfi [28] highlighted the impact of introducing context on the automatic correction of spelling errors. They proposed a mechanism that allows to exploit thematic contextual information improving the accuracy of the spelling correction system. This study aims to solve the problem encountered at the level of automatic spelling correction systems, which lies in the classification of the desired solution in the last position of the list of given suggestions. The proposed correction system is based on a dictionary, which contains a probability distribution of occurrence of a word in various contexts.

Li et al. [29] addressed the spelling correction problem at the word level, by correcting the spelling of each token regardless of additional token insertion or deletion. They proposed a context-aware stand-alone neural spelling correction where they utilized both spelling information and global context representations to detect and correct misspellings in the form of a sequence labeling task by fine-turning a pre-trained language model. They enhanced the state-of-the-art results by 12.8% precision score.

Siklosi et al. [30] proposed a method for the automated correction of spelling mistakes in Hungarian clinical records. A word-based algorithm is applied to generate a list of suggested corrections for each erroneous word. Then, the spelling correction problem is represented as a translation task, where the source language is the incorrect text and the target language is the corrected text. A statistical machine translation (SMT) decoder executes the error correction task.

Wang et al. [31] proposed a novel contextual method by adding a light-weight contextual spelling correction model on top of transducer-based automatic speech recognition (ASR) systems. They introduced the context into the spelling correction model with a shared context encoder that encodes context expressions into hidden embeddings. A filtering algorithm is used to deal with large-size context lists. The results outperform the baseline method by reducing approximately 50% in relative word error rate.

The previously cited works presented considerable results in automatic spelling correction. However, no work has taken into account all of these errors by exploiting the strong semantic links between the distant words in the considered context, including the possible correction candidates of these errors in order to improve their ranking and prioritize the desired solutions.

3. GLOBAL CONTEXTUAL SPELLING CORRECTION

The approaches used in spelling correction, including those using context, only consider correct words near the error to be corrected, thus ignoring other erroneous words in the same context. These methods are referred to as single methods. The consideration of the error-word correction candidates preceding and following the target error can refine the list of suggestions of the latter and improve the rank of the desired solution. This method has never been implemented before.

Definition: Let $ph = w_1 w_2 \dots w_T$ be a given sentence that contains a set of erroneous words in different positions. The global spelling correction, is the operation which consists in first returning all the lists of corrections proposed for each single error (Table I), then identifying the set of solutions of all the erroneous words which are the most appropriate to the global context of the sentence ph. The global spelling correction is done in three stages:

• The detection of erroneous words, by referring to the vocabulary of the system.

• Generating lists of corrections associated with each erroneous word, using one of the spelling correction methods

(in our case, we used the edit distance). The solutions are classified in a list according to the value of this distance.

• By taking into account the global context of the sentence, we identify among these lists, the solutions (associated with all the erroneous words in the sentence) the most suitable with this context. The context is introduced via language models. The Viterbi algorithm [32] ensures the coherence of the global correction by determining the most probable path built from the correct words and the correction candidates of the errors.

Example: Let us consider the following Arabic sentence:

عجق الوجيد إلى القدرس وحضر حفضة اللغة الفرنسية.

This sentence contains the misspellings: عجق, الوجيد, القدرس, حفضة. The generated lists of solutions for these four erroneous words, classified in ascending order of the edit distance, are:

List(عجق) = { عرق,عنق,علق,عاد }

List(الوجيد) = { الوريد,الوصيد,الوليد,الولد }

List(القدرس) = {القدس,القدر,المدرس,المدرسة}

List(حفضة) = {حفلة,حصيد,حصاد,حصة}

If we consider the error-by-error correction method (Table I), taking into account the context by using the edit distance and language models, we risk having the following solutions with the maximum probability:

• عرق الوريد إلى القدس وحضر حفلة اللغة الفرنسية

• عنق الوليد إلى القدس وحضر حفلة اللغة الفرنسية

The first sentence is the result of selecting the first solutions from the respective lists of suggestions obtained with the edit distance. The second sentence is obtained with the language models to find the most probable words. Both of these solutions are semantically unwanted. To remedy this problem, we propose a method that takes into account the global context of the sentence by taking into consideration the semantic links between the following words: {عاد، الولد، المدرسة، حصة}. Our method gives the following solution:

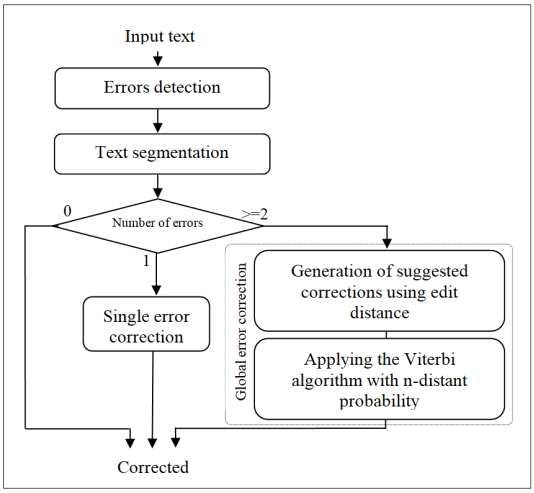

Figure 1. Process of correcting spelling errors in a text

عاد الولد إلى المدرسة وحضر حصة اللغة الفرنسية

4. PROPOSED METHOD

In this paper, we propose a new method of contextual spelling correction, which allows to globally and simultaneously correcting spelling mistakes in a text containing more than one spelling error.

A. Modeling the problem

Our method first involves detecting spelling errors in the target text and proposing for each detected error a set of candidate words generated using the edit distance metric. The number of correction suggestions is the same for all errors. Next, a context-related metric is defined for each suggestion. This quantity is calculated with the n-distant language model [14] [27] and the edit distance. Finally, by applying the Viterbi algorithm [32], we get a contextual spelling correction by selecting the optimal path in terms of probability, composed of the initial correct words and the suggestions of the detected errors. The process of correcting spelling errors in a text can be generally schematized by the scheme in Fig. 1.

In the rest of this paper, we will consider $ph = o_1 o_2 \dots o_T$ a given sentence containing a set of erroneous words and $V = \{w_1, w_2 \dots w_N\}$ is the vocabulary of our system. The first stage of our method consists in detecting misspelled words by checking the existence of each word of the input sentence in the vocabulary of our system. Then, for each detected erroneous word, we look up in the vocabulary for the closest correct words and then classify them by decreasing edit distance. Let m be an integer. For each erroneous word o_{ri}, we consider the set of the first m suggestions $(w_{ri}^1, w_{ri}^2, ...w_{ri}^m)$ generated using the edit distance. The following diagram in Fig. 2 represents the sequence of words ph and the suggested corrections for each misspelled word. Our method aims to select, for each erroneous word, the correct suggestion among those

TABLE I. Contextual correction of several errors

عجق	الوجيد	إلى	القدرس	وحضر	حفضة	اللغة الفرنسية
عرق	الوريد		القدس		حفلة	
عنق	الوصيد		القدر		حصيد	
علق	الوليد		المدرس		حصاد	
عاد	الولد		المدرسة		حصة	

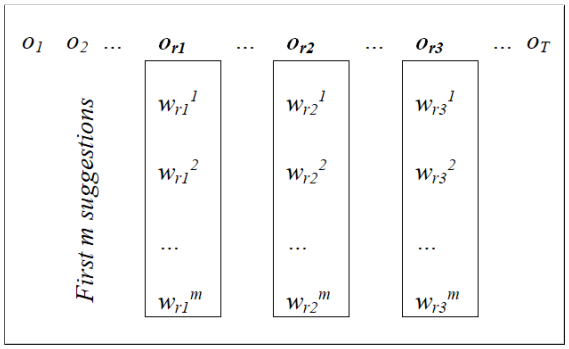

Figure 2. Representation of the correction suggestions

generated by the edit distance taking into account the global context.

B. n-gram Language Models

We consider a sequence of k words $S_k = w_1, w_2 \dots w_k$. Probabilistic language models consist in assigning a probability to S_k. An n-gram language model is a model derived from information theory [33]. It only considers the last $n-1$ words before the target word. The n-gram model verifies (1):

$$Pr(w_i/w_1, \dots w_{i-1}) = Pr(w_i/w_{i-(n-1)}, \dots w_{i-1}) \quad (1)$$

The complexity of these models increases when n increases. The most used model for spelling correction is the bigram model; for its simplicity. The main disadvantage of this last model is to favour corrections relating to the word immediately preceding the erroneous word. Whereas, strong links can exist between distant words in a context. Two language models, inspired by n-gram models, are proposed to overcome this drawback [15]: n-distant-max and n-distant-avg models.

1) n-distant-max Model: The n-distant-max model assumes that the probability of occurrence of a word w_i after a sequence $w_1, w_2, \dots w_{i-1}$ can be satisfactorily given by the maximum of the distant bigram probabilities based on the $n-1$ previous observations. This probability verifies (2). The estimation of these probabilities is given by the maximum likelihood method in (3). $O(w_j)$ is the number of occurrences of the word w_j in the training corpus and $O((w_j w_i)/context)$ is the number of times the word w_j appears before the word w_i in a context of size n.

$$Pr(w_i/w_1, \dots w_{i-1}) \approx \max_{i-n+1 \le j \le i-1} Pr(w_i/w_j) \quad (2)$$

$$Pr(w_i/w_j) = O(w_j w_i/context) \Big/ O(w_j) \quad (3)$$

2) n-distant-avg Model: The n-distant-avg model assumes that the probability of occurrence of a word w_i after a sequence $w_1, w_2 \dots w_{i-1}$ can be satisfactorily given by the average of the bigram probabilities based on the $n-1$ previous observations. This probability is given by (4). The estimation of these probabilities is given by the maximum likelihood method.

$$Pr(w_i/w_1, \dots w_{i-1}) \approx (1/(n-1)) \sum_{j=i-n+1}^{i-1} Pr(w_i/w_j) \quad (4)$$

C. Defining a Hidden Markov Model

To be able to globally correcting all the existing errors in the sentence" $ph = o_1 o_2 \dots o_T$, we seek to find the optimal path, which optimizes the probability given by (5). To facilitate the calculation of this probability, we use a Hidden Markov Model (HMM) of order n. This model is defined by the two processes X_t and Y_t [34]. $V = \{w_1, w_2 \dots w_N\}$ is the set of hidden states of our model, which is the vocabulary of the system. We suppose that, for each detected error, the set of the first m suggestions is significantly representative.

$$\max_{w_{i_1}, \dots w_{i_T} \in V} Pr(o_1, o_2, \dots o_T/w_{i_1}, w_{i_2}, \dots w_{i_T}) \quad (5)$$

• X_t is a stochastic process with values in V. This process is a HMM of order n, and it satisfies equation (6). Since the context of a word does not only depend on the previous word, and to increase the efficiency of our model we considered the HMM of order n, not of order 1.

$$\begin{aligned} &Pr(X_t = w_j | X_{t-1} = w_i, \dots X_1 = w_{i_1}) \\ &\quad = Pr(X_t = w_j | X_{t-1} = w_i, \dots X_{t-n} = w_{i_{t-n}}) = a_{.j} \end{aligned} \quad (6)$$

• Y_t is an observable stochastic process with value in the set ph. Y_t verifies:

$$\begin{aligned} &Pr(Y_t = o_t | X_t = w_j, \dots X_1 = w_{i_1}, Y_{t-1} = o_{t-1}, \dots Y_1 = o_1) \\ &\quad = Pr(Y_t = o_t | X_t = w_j) = b_j(o_t) \end{aligned} \quad (7)$$

To solve this problem, we use the Viterbi algorithm. For this, we define:

$$\delta_t(w_j) = \max_{w_{i_1},\ldots w_{i_{t-1}} \in V} Pr(o_1, o_2, \ldots o_t / w_{i_1}, w_{i_2}, \ldots w_{i_{t-1}}) \quad (8)$$

This gives us Viterbi's recurrent formula [32], with considering only the first m suggestions:

$$\delta_t(w_j) = \max_{1 \le i \le m} \delta_{t-1}(w_i) * a_{.j} * b_j(o_t) \quad (9)$$

With $t = 1, \ldots T$ and $j = 1, \ldots m$. To simplify the calculations, we apply the logarithm to this formula, which is an increasing function, this equation becomes:

$$\delta_t(w_j) = \max_{1 \le i \le m} \{\delta_{t-1}(w_i) + a_{.j} + b_j(o_t)\} \quad (10)$$

1) Estimation of the Parameters of our HMM Model: The model we have defined requires the estimation of the following parameters.

- π_j is the initial probability of the current state w_j, for all the vocabulary words V. π_j indicates the probability that the word w_j is at the beginning of the sentence. It is estimated by (11).

- $a_{.j}$ is the transition probability from previous state w_i in the window of size n preceding the current state w_j, for all the vocabulary words V.

- $b_j(o_t)$ is the state observation likelihood of the observation o_t given the current state w_j, for all the vocabulary words and for all words $o_1, o_2, \ldots o_T$.

$$\pi_j = Pr(w_j/.) = O(w_j/.) \Big/ O(w_j) \quad (11)$$

$O(w_j/.)$ is the number of occurrences that the word w_j is at the beginning of the sentence, and $O(w_j)$ is the number of occurrences of the word w_j. In a previous paper [14], we developed the two models n-distant-max and n-distant-avg to estimate the probabilities of the n-gram language model. The use of the n-distant-max model makes it possible to estimate the $a_{(.j)}$ by (12).

$$a_{.j} = \max\{Pr(X_t = w_j | X_{t-1} = w_i), \ldots Pr(X_t = w_j | X_{t-n} = w_{i_{t-n}})\} \quad (12)$$

The n-distant-avg model uses the following formula to estimate $a_{.j}$:

$$a_{.j} = \sum_{k=1}^{n} Pr(X_t = w_j | X_{t-k} = w_{i_{t-k}}) \quad (13)$$

For the estimate of the probabilities $b_j(o_t)$, it is given by (14). $D_L(o_t, w_j)$ is the Damerau-Levenshtein distance between the two words o_t and w_j.

$$b_j(o_t) = 1 \Big/ [1 + (D_L(o_t, w_j) + 2)^2] \quad (14)$$

2) Algorithm for Calculating the Optimal Path: To find the global solution, we calculate the optimal path based on the Viterbi algorithm described in Fig. 3 below. Given the observation sequence ph and having estimated our HMM parameters, we identify the best suggestions of correction of all the misspelled words in ph. The input and the output for the Viterbi algorithm are as follows:

```
Function Viterbi_decoding(ph , (Sk), π, A, B)
//initialisation step: t=1
For state j in 1... m do
    Viterbi_decoding [j, 1]  ← π_j * b_j(o_1)
    backpointer [j, 1]  ← 0
End For
//recursion step
For time t in 2... T do
    For state j in 1... m do
        Viterbi_decoding [j, t] ←
            max_{i∈[1,m]} {Viterbi_decoding [i, t − 1] + a_ij + b_j(o_t)}
        backpointer [j, t] ←
            argmax_{i∈[1,m]} {Viterbi_decoding [i, t − 1] + a_ij + b_j(o_t)}
    End For
End For
//termination step
max_path_probability  ← max_{j∈[1,m]} Viterbi_decoding [j, T]
best_path_pointer  ← argmax_{j∈[1,m]} Viterbi_decoding [j, T]
best_path  ← the path starting at the state
best_path_pointer, and follows backpointer[]
to states back in time t.
return max_path_probability, best_path
```

Figure 3. Viterbi algorithm for calculating the optimal path.

Input:

- A sequence of observations $ph = (o_1, o_2 \ldots o_T)$
- The state spaces $S_k = \{w_{k_1}, w_{k_2} \ldots w_{k_m}\}$
- An array of initial probabilities $\pi = (\pi_1, \pi_2 \ldots \pi_m)$
- The transition matrix $A = (a_{.j})$ of size TxT
- The emission matrix $B = (b_j(o_t))$ of size Txm

Output:

- The most likely hidden state sequence $best\_path = (s_1, s_2 \ldots s_T)$
- $max\_path\_probability$.

5. EXPERIMENTS AND RESULTS

In this section, we will present the implementation steps of our approach as well as the results obtained. The first step is to choose a training corpus for the estimation of the parameters of our HMM. Then, we present the test corpus from which we generate a set of erroneous words based on the editing operations: deletion, insertion, substitution and transposition. Finally, we draw up a performance comparison of our correction method with the most widely used methods in this direction.

TABLE II. Number of entries by dictionary

Window size	Number of entries
n=2	3,206,353
n=6	17,796,407
n=7	21,101,418

A. Pre-processing and Learning

For the evaluation of our method, we used the "Kalimat" corpus [35] to train our HMM. This corpus is a set of multidisciplinary documents covering a large number of vocabularies of the Arabic language in different fields. It contains approximately 20,291 articles and covers six topics. For better use of this corpus, we proceeded to clean it of punctuation marks, numbers, and characters from other languages... Then, we extracted its vocabulary and generated bigram and n-distant bigram transition dictionaries for different window sizes up to $n = 10$. After checking, we retained the values $n = 6$ and $n = 7$ as being the values achieving the best correction accuracy [15]. The dictionaries contain transition probabilities (Table II). For transitions not represented in these dictionaries, a very low probability is assigned (to smooth the values). We used a part of the corpus for the test. We constructed 1879 sequences each containing 30 words. A copy of this test corpus was dedicated to the generation of spelling errors. Each sequence contains randomly 3, 4 or 5 erroneous words. In the creation of errors, we randomly chose the target words to create different types of editing errors (deletion, insertion, substitution and reversal). The targeted words are not very short (three or more letters). To keep a minimum of resemblance with the original word, a limit of three editing operations is set as a maximum error to be made on each word. In total we have 5642 erroneous words.

B. Evaluation

The development of automatic correction systems requires a method of evaluating the results obtained and comparing these results with other systems. The correction operation often depends on the language used and its scope. Thus, there is no generic method of performance comparison. In this paper, we use the accuracy measurement as the most commonly used evaluation method and the most suitable for our case. Table III illustrates the correction accuracy, for the first given suggestion, corresponding to the single and global correction methods using the edit distance and the language models: Bigram, 6-distant-avg, 7-distant-avg, 6-distant-max and 7-distant-max.

We find that, for the first correction suggestion of the single method, the best correction method is the one using the 6-distant-avg language model with a correction accuracy of 79.58%. On Table III, we notice that, for the first correction suggestion of the global methods, the best method of correction is the one using the 6-distant-avg language model with an accuracy of 93.69%. Table IV presents a comparison of the correction accuracies, for the first position, between the best single method and the best global method.

On Table IV, we can see that the global correction method using the 6-distant-avg language model is the best correction method with a difference of 14.11% accuracy compared to the best single correction method with the same language model. This clearly shows the advantage of the global method in correcting spelling errors.

C. Comparison with two spell checkers

In addition to the evaluation of our method, we have drawn up a comparison between the correction results obtained by two spell checkers and our method, on a sample of 34 sentences each containing 2 to 4 erroneous words, with a total of 103 errors. The spelling correctors used are: Google Docs and ArabicCorrection (www.arabiccorrection.com). Table V presents the obtained results.

The global method achieves a correction rate of 97.09% against 52,43% for Google Docs and 10.68% for ArabicCorrection. Note that Google Docs offers a suggested correction by clicking on the erroneous word while ArabicCorrection allow proposing to the user, for each detected error, a list of correction candidates. As an example, we consider the following sentence which contains 4 misspelled words (فسياؤ، سخرجا، صوبتا، اختصلاها):

السلبية التي قد تشكل ضغطا فسياؤ عليها وبالتالي فإن هذه الضغوط تجد لها سخرجا قد يساهم في تحسين العلاقات الأسرية إن صوبتا المرأة العاملة التي تظهرها الوسائل الإعلامية على اختصلاها

The correction given by our global method:

السلبية التي قد تشكل ضغطا نفسيا عليها وبالتالي فإن هذه الضغوط تجد لها مخرجا قد يساهم في تحسين العلاقات الأسرية إن صورة المرأة العاملة التي تظهرها الوسائل الإعلامية على اختلافها

The correction obtained for the global method corresponds to the desired correction. The correction given by the corrector of Google Docs, is as follows:

السلبية التي قد تشكل ضغطا فيزياء عليها وبالتالي فإن هذه الضغوط تجد لها مخرجا قد يساهم في تحسين العلاقات الأسرية إن صوبها المرأة العاملة التي تظهرها الوسائل الإعلامية على اختصارها

TABLE III. Correction accuracy (%) for single and global methods

	Edit distance	Bigram	6-distant-avg	7-distant-avg	6-distant-max	7-distant-max
Single	24,16	67,09	79,58	79,14	78,06	77,9
Global	24,16	89,04	93,69	93,51	93,48	93,12

TABLE IV. Accuracy (%) of the best single and the best global correction methods

	Single	Global
Language model	6-distant-avg	6-distant-avg
Accuracy	79,58	93,69

TABLE V. Comparison of correction results

	G. docs	Global	ArCorr
# corrected errors	54	100	11
Accuracy (%)	52,43	97,09	10,68

We note that two accepted corrections are provided for this sentence by the corrector of Google Docs. For ArabicCorrection, the correction obtained by considering the first suggestion of the correction lists is the following:

السلبية التي قد تشكل ضغطا فسياء عليها وبالتالي فان هذه الضغوط تجد لها يخرجا قد يساهم في تحسين العلاقات الأسرية إن صوبها المرأة العاملة التي تظهرها الوسائل الإعلامية على اختصاها

For this sentence, ArabicCorrection does not give any desired correction in the first position of the list of suggested corrections. In this comparison sample, it is seen that our method can provide the desired correction suggestion as a first choice even for misspellings that have undergone more than one editing operation.

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented an effective method able to correct spelling errors in Arabic textual documents. It makes it possible to identify and correct globally and simultaneously several errors in a given context. This method is inspired by several observed examples of sentences containing spelling errors in which the desired correction necessarily depends on the correction of other errors. In most cases, the potential correction of each error is strongly correlated to the words related to the given context. The method uses n-distant bigram language models and HMMs to globally correct spellings in its specific context. The estimation of the model's parameters was carried out on an Arabic multipurpose corpus. Thus, the method has been applied to the Arabic language, but it can be adapted and applied to other languages. To assess the effectiveness of the proposed method, we have conducted a set of experiments on a test corpus. The results obtained are better than those of the error-by-error correction method. The accuracy reaches a rate of 93.69% for the solutions found in the first position of the list of solutions proposed for each erroneous word. The high-level correction accuracy of our method allows its integration in many applications such as OCR and speech recognition. As future work, the learning corpus can be expanded to enlarge the representativeness of words and further increase the accuracy of correction. Our method can also be completed by including real-word errors using the same language models. The consideration of the syntactic layer in spelling correction is a third perspective of this work.

REFERENCES

[1] K. M. Zakaria, "Natural language processing and computational linguistics 2: semantics, discourse and applications," *John Wiley & Sons*, vol. 2, 2017.

[2] A. S. S. N. Srihari and S. W. Lam, "Optical character recognition (ocr)," *Encyclopedia of Computer Science*, pp. 1326–1333, 2003.

[3] C. C. R. Dabre and A. Kunchukuttan, "A survey of multilingual neural machine translation," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–38, 2020.

[4] X. W. D. Wang and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, p. 1018, 2019.

[5] J. Haider and O. Sundin, "Invisible search and online search engines: The ubiquity of search in everyday life," *Taylor & Francis*, 2019.

[6] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1–22, 2009.

[7] G. Julian, "What are the most spoken languages in the world," *Retrieved May*, vol. 31, p. 2020, 2020.

[8] C. Liensberger, "Context sensitive auto-correction," *U.S. Patent 9,218,333*, vol. issued December 22, 2015.

[9] R. A. E. W. V.W. Berninger, K.H. Nielsen and W. Raskind, "Writing problems in developmental dyslexia: Under-recognized and under-treated," *Journal of school psychology*, vol. 46, no. 1, pp. 1–21, 2008.

[10] G. H. C.A. MacArthur, S. Graham and S. DeLaPaz, "Spelling checkers and students with learning disabilities: Performance comparisons and impact on spelling," *The Journal of Special Education*, vol. 30, no. 1, pp. 35–57, 1996.

[11] G. K. D.J. Montgomery and M. Coutinho, "The effectiveness of word processor spell checker programs to produce target words for misspellings generated by students with learning disabilities," *Journal of Special Education Technology*, vol. 16, no. 2, pp. 27–42, 2001.

[12] K. Kukich, "Techniques for automatically correcting words in text," *Acm Computing Surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.

[13] R. M. E. Brill, "An improved error model for noisy channel spelling correction," *Proc. 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293, 2000.

[14] S. L. H. G. S.L. Aouragh, A. Yousfi and M. Nejja, "A new estimate of the n-gram language model," *Procedia Computer Science*, vol. 189, pp. 211–215, 2021.

[15] A. Y. M. N. H. G. S. Laaroussi, S.L. Aouragh and S. O. E. Alaoui, "New language models for spelling correction," *International Arab Journal of Information Technology*, vol. 19, no. 6, pp. 942–948, 2022.

[16] S. Deorowicz and M. G. Ciura, "Correcting spelling errors by modelling their causes," *International journal of applied mathematics and computer science*, vol. 15, no. 2, pp. 275–285, 2005.

[17] A. Yunus and M. Masum, "A context free spell correction method using supervised machine learning algorithms," *International Journal of Computer Applications*, vol. 975, p. 8887, 2020.

[18] F. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of the ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[19] W. G. M.D. Kernighan, K. Church, "A spelling correction program based on a noisy channel model," *Proc. 13th Conf. on Computational Linguistics*, pp. 205–210, 1990.

[20] A. Alwabel, "Coedit: A novel error correction mechanism in compilers using spelling correction algorithms," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8490–8500, 2021.

[21] M. Elmi, "A natural language parser with interleaved spelling correction supporting lexical functional grammar and ill-formed input," *Ph.D. thesis, Illinois Institute of Technology*, 1994.

[22] J. Zobel and P. Dart, "Phonetic string matching: Lessons from information retrieval," *Proc. 19th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, ACM Press*, pp. 166–172, 1996.

[23] O. Büyük, "Context-dependent sequence-to-sequence turkish spelling correction," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 4, pp. 1–16, 2020.

[24] K. Church and W. Gale, "Probability scoring for spelling correction," *Statistics and Computing*, vol. 1, no. 2, pp. 93–103, 1991.

[25] H. G. S.L. Aouragh and A. Yousfi, "Adaptating the levenshtein distance to contextual spelling correction," *International Journal of Computer Science and Applications*, vol. 12, no. 1, pp. 127–133, 2015.

[26] Y. Y. R. Dong and T. Jiang, "Spelling correction of non-word errors in uyghur–chinese machine translation," *Information*, vol. 10, no. 6, p. 202, 2019.

[27] S. A. S. Laaroussi and A. Yousfi, "Distant n-gram language model for contextual spelling correction applied to arabic language," *Proc. 2nd international conf. on Embedded Systems and Artificial Intelligence*, 2021.

[28] M. Nejja and A. Yousfi, "Context's impact on the automatic spelling correction," *International Journal of Artificial Intelligence and Soft Computing*, vol. 6, no. 1, pp. 56–74, 2017.

[29] H. L. X. Li and L. Huang, "Context-aware stand-alone neural spelling correction," *arXiv preprint arXiv:2011.06642*, 2020.

[30] A. N. B. Siklósi and G. Prószéky, "Context-aware correction of spelling errors in hungarian medical documents," *Computer Speech & Language*, vol. 35, pp. 219–233, 2016.

[31] S. Z. X. Wang, Y. Liu and J. Li, "A light-weight contextual spelling correction model for customizing transducer-based speech recognition systems," *arXiv preprint arXiv:2108.07493*, 2021.

[32] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[33] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. of IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[34] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[35] M. El-Haj and R. Koulali, "Kalimat a multipurpose arabic corpus," *Second workshop on Arabic corpus linguistics (WACL-2)*, pp. 22–25, 2013.

Saida Laaroussi is currently PhD student in the ES-Lab in ENSA, at Ibn Tofail University in Kenitra, Morocco. She received her engineering degree in Computer Science from the ENSIAS at Mohamed V University in Rabat, Morocco, in 2010. Her main research interests include Machine Learning and Natural Language Processing.

Abdellah Yousfi is Professor at the Faculty of Law, Economics and Social Sciences of Souissi at Mohamed V University in Rabat. He is member of the ICES Team in the ENSIAS, at Mohamed V University in Rabat, Morocco. His research interests include creation of corpora for the Arabic language, Arabic speech recognition, Arabic handwriting recognition and correction of Arabic spelling errors. He is reviewer of several journal such as Journal of King Saud University, Computer and Information Sciences, Egyptian Informatics Journal.

Si Lhoussain Aouragh is permanent qualified professor in the ENSIAS at Mohamed V University in Rabat. He is president of the Association of Arabic Language Engineering in Morocco, and member of several scientific research associations in Morocco. Member of several research teams and laboratories. His main research interests include Computational Linguistics, Artificial Intelligence, Machine Learning, Natural Language Processing.

Said Ouatik El Alaoui is working as Professor of Computer Science in the ENSA, Kenitra where he is currently the head of the ES-Lab at Ibn Tofail University, Morocco. His research interests include Machine and Deep Learning and their applications, Natural Language Processing, Information Retrieval, Text summarization, Biomedical Question Answering, Biomedical Information Extraction, and Arabic Document Clustering and Categorization, High-dimensional indexing and Content-Based Image Retrieval.