New Language Models for Spelling Correction

Saida Laaroussi
IT, Logistics and Mathematics, Ibn Tofail University, Morocco
saida.laaroussi1@uit.ac.ma

Si Lhoussain Aouragh
IT and Decision Support System, Mohamed V University, Morocco
aouragh@hotmail.com

Abdellah Yousfi
Department of Economics and Management, Mohamed V University, Morocco
yousfi240ma@yahoo.fr

Mohamed Nejja
Department of Software Engineering, Mohamed V University, Morocco
mohammed.nejja@gmail.com

Hicham Geddah
Department of Computer Science, Mohamed V University, Morocco
h.gueddah@um5s.net.ma

Said Ouatik El Alaoui
IT, Logistics and Mathematics, Ibn Tofail University, Morocco
s_ouatik@yahoo.com

Abstract: *Correcting spelling errors based on the context is a fairly significant problem in Natural Language Processing (NLP) applications. The majority of the work carried out to introduce the context into the process of spelling correction uses the n-gram language models. However, these models fail in several cases to give adequate probabilities for the suggested solutions of a misspelled word in a given context. To resolve this issue, we propose two new language models inspired by stochastic language models combined with edit distance. A first phase consists in finding the words of the lexicon orthographically close to the erroneous word and a second phase consists in ranking and limiting these suggestions. We have applied the new approach to Arabic language taking into account its specificity of having strong contextual connections between distant words in a sentence. To evaluate our approach, we have developed textual data processing applications, namely the extraction of distant transition dictionaries. The correction accuracy obtained exceeds 98% for the first 10 suggestions. Our approach has the advantage of simplifying the parameters to be estimated with a higher correction accuracy compared to n-gram language models. Hence the need to use such an approach.*

Keywords: *Spelling correction, contextual correction, n-gram language models, edit distance, NLP.*

Received January 1, 2021; accepted January 19, 2022
https://doi.org/10.34028/iajit/19/6/12

1. Introduction

Spelling correction is one of the oldest Natural Language Processing (NLP) applications. It consists of suggesting one or more words closest to a misspelled word. Its quality depends on lexical resources and algorithms used. Generally, a distinction is made between two types of spelling correction: correction out of context and correction based on context. The principle of context-dependent type is to correct the misspelled word taking into account the neighboring words. Whereas the context-free correction consists in suggesting, after the detection of the misspelled word, one or more words regardless of the context. Several works and techniques have been developed to correct spelling errors out of the context, among which we cite:

- The technique based on the edit distance [3, 13]. It calculates the number of editing operations to transform a word into another word.
- The Hidden Markov Model Technique (HMM) [2].
- The technique based on finite state automata [1, 17].
- The Alpha-code technique [20] which uses a coded representation for a given word.
- The influence of the similarity and proximity of keyboard keys for Arabic characters on editing errors, and the introduction of the concept of morphological analysis in the correction phase [7, 8, 10, 14, 16].

In this paper, a context-dependent approach is proposed for spelling correction. In most of the work in this area, the methods used do not perfectly correct these errors; in some cases they offer solutions which are very far from the context of the text. To examine the problem, we first created several types of spelling errors in Arabic sentences. Then, we used Levenshtein algorithm to correct them without taking into account their context. A manual study on those errors and on suggested solutions allowed us to develop two new language models based on n-gram language models combined with Levenshtein distance. These models have significantly enhanced the correction efficiency.

The rest of this paper is organized as follows. First, we present the related work in section 2. Section 3 underlines the importance of contextual correction. Section 4 presents the n-gram language models. We introduce our new language models in section 5. Section 6 is devoted to the combination of the new language models and Levenshtein algorithm. Section 7 concerns the implementation phase. Finally, section 8 outlines the conclusion and the future work.

2. Related Work

Considering context in NLP applications has improved expected results [9]. For spelling correction, the context was introduced to overcome the problem of hidden (real-word) errors which are lexically valid but are semantically far from the context. For this kind of errors, a very famous method in contextual correction is often used. It is inspired by the cosine similarity [18].

Context-free correction is the first phase of spelling correction. Then, taking into account the context of the misspelled word, the proposals returned in the context–free correction are limited and re-ranked.

A minority of methods and techniques have been developed in order to introduce the contextual level in the correction process. Among these methods we quote:

- Jones and Martin [12] proposed a method based on the latent semantic analysis of the text in order to improve the correction and also limit the number of suggested solutions.
- Sharma and Gupta [19] used the trigram language model and another Bayesian approach to correct spelling mistakes made in their input contexts.
- Nejja and Yousfi [15] proposed a correction approach which combines between a metric method of lexical similarity via Levenshtein distance and a method based on context, through a well-defined training corpus made up of documents collected on the Internet.
- Farra *et al*. [4] presented a general discriminant model for correcting spelling errors in Arabic. Unlike previous work, the proposed approach teaches how to correct a variety of error types without being guided by manually selected constraints or language-specific functionality. The model is applied to correct text errors in Egyptian Arabic dialect, achieving a 65% reduction in the error rate on words compared to the entry baseline and improving compared to the prior advanced system.
- Golding and Roth [6] presented an algorithm combining variants of Winnow and weighted majority voting. This algorithm is able to recognize approximately 96% of contextual spelling mistakes, in addition to ordinary non-verbal spelling mistakes, introducing the syntactic level in the correction phase.
- Fossati and Eugenio [5] addressed the problem of spell checking of real words. They propose a methodology based on a mixed trigram language model. The approach was evaluated in terms of success rate, false positive rate and coverage. Experiments show promising results in terms of detection and correction success rates, although the false positive rate is still high.

Nevertheless, the English and French speaking communities are more advanced in the field of automatic spelling correction compared to the Arab ones, who suffer from the insufficiency of standardized and freely usable resources.

3. Spelling Correction Based on the Context

Automatic spelling correction is a discipline that closely associates linguists and computer scientists. In general, spelling errors often need to be corrected taking into account the context of misspelled word. This type of correction is done in three stages:

- *Verification*: check whether an input word is correct or not, i.e., present or not in the lexicon.
- *Context-free correction*: propose the words closest lexically to the misspelled word.
- *Context-dependent correction*: select among the given solutions those most appropriate to the context of the misspelled word.

Example:

قام الولد بتصحيح أخطائه الإصلاخية خلال حصة الإملاء

For the misspelled word الإصلاخية, if we don’t consider its context, the word الإصلاحية for example, is most likely close compared to the word الإملائية with a Levenshtein distance of 1 against 2. Whereas, if we take into account the context, the solution الإملائية is the most probable solution compared to other suggestions, in particular, الإصلاحية.

Several research works (see second section) have been conducted in order to introduce the context in the correction phase. A fairly large number of these works use the n-gram language models. In the following section, we give a little reminder of these models and how they are used in context-sensitive spelling correction.

4. N-Gram Language Models

Let $S^k = w_1, \dots w_{i-1}, w_{err} \dots w_k$ be a given sentence where w_{err} is a wrong word. Probabilistic language models consist in assigning a probability to a sequence of words. An n-gram language model is a model derived from information theory [11]. It takes into account only the last *n-1* words before the target word. The n-gram model verifies the following equation:

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

For n=3, the model is called a trigram. For n=2, the model is called a bigram. In general, almost all of the work that has dealt with spelling errors in context has used bigram language model. The main disadvantage of this model is that it gives more priority to corrections which are related to the word just before the misspelled word. But often, the word may have stronger connections to words that are farther apart than the word preceding it, as in the following example:

يوجد في هذه الصحيفة الكثير من الأخطاس

For the misspelled word الأخطاس, the context-free correction phase gives several solutions الأقطار, الأخماس, الأخطاء, الأخطار. If we consider the bigram language model, all these words are related to the word من which is just before the misspelled word الأخطاس. But the meant word الأخطاء, which is most correct in the global context, is related to the word الصحيفة which is further from the misspelled word. Without taking this word into account in the correction, we never get to the right solution. To remedy to this drawback, we propose two new language models: *n-distant-max* and *n-distant-avg*.

5. Introducing New Language Models

We consider a history of size n.

5.1. N-Distant-Max Model

To calculate the probabilities of Equation (1), the n-distant-max model assumes that the probability of appearance of a word w_i after a sequence $w_1,\ldots w_{i-1}$ can be satisfactorily given by the maximum of the bigram distant probabilities based on the *n-1* previous observations. This probability is given by:

$$\Pr(w_i / w_1,\ldots w_{i-1}) = \max_{i-n+1 \le j \le i-1} \Pr(w_i / w_j) \qquad (2)$$

Where $Pr(w_i/w_j)$ is the distant bigram probability of w_i knowing w_j. This approximation has two advantages over the n-gram model. First, it requires fewer parameters to estimate. Second, certain words in the history of a given word provide no information about it. Therefore, they make the probability of the n-gram very small compared to the n-distant-max model. The estimation of the parameters of this new model is given by the following equation:

$$\Pr(w_i / w_j) = \frac{O(w_j w_i / context)}{O(w_j)} \qquad (3)$$

Where $O(w_j)$ is the number of occurrences of the word w_j in the learning corpus and $O(w_j w_i/context)$ is the number of times the word w_j occurs before the word w_i in a context of size n. For a given context, the greater the number of different words that can complete it, the smaller is the probability of each word. Conversely, if there is only one word capable of completing a given context, the probability of its occurrence becomes 1.

For this model, we notice that there are always solutions that are misclassified in returned solutions list. By analyzing examples of these solutions, we have deduced that this problem amounts to the fact that the n-distant-max model takes into account only one word which is most closely related to the proposed solution. However, there are cases where the words are linked to more than one word, as shown in the following example:

دخل التلميذ إلى المدرسة وهو يحمل المحفثة

The word المحفظة, which is the meant correct word for the misspelled word المحفثة, is related to the three words يحمل, التلميذ and المدرسة. To take this remark into account, we propose a second model called the n-distant-avg model.

5.2. N-Distant-Avg Model

To calculate Equation (1) probabilities, the n-distant-avg model assumes that the probability of appearance of a word w_i after a sequence $w_1,\ldots w_{i-1}$ can be satisfactorily given by the average of bigram probabilities based on the *n-1* previous observations. This probability realizes:

$$\Pr(w_i / w_1,\ldots w_{i-1}) = \frac{1}{n-1} \sum_{j=i-n+1}^{i-1} \Pr(w_i / w_j) \qquad (4)$$

6. Combining the New Language Models with the Levenshtein Algorithm

6.1. Levenshtein Algorithm

Let be a misspelled word $w_{err}=e_1e_2\ldots e_n$ and $w_i=c_1c_2\ldots c_m$ a word from the lexicon. Every character e_i has for index *i* and every character e_j has for index *j*. Levenshtein distance between these two words is given by the following recurring equation:

$$D_L(i,j) = \begin{cases} D_L(0,0)=0 \\ \min\{D_L(i-1,j)+1; D_L(i-1,j-1)+sub(e_i,c_j); D_L(i,j-1)+1\} \end{cases}$$
$$sub(e_i,c_j) = \begin{cases} 0 & if \quad e_i = c_j \\ 1 & if \quad e_i \neq c_j \end{cases} \qquad (5)$$

Levenshtein distance between the two words w_{err} and w_i noted $D_L(w_{err}, w_i)$ is obtained by $D_L(n,m)$.

6.2. Introduction of the N-Distant-Max Language Model in the Levenshtein Algorithm

Let $w_1w_2\ldots w_{i-1}w_{err}$ be a context containing the misspelled word w_{err}, and V={v1,v2,…vM} be the system's vocabulary. To be noted that $D_L(w_{err}, w_k)$ is the Levenshtein distance between w_{err} and w_k. To find the right suggestion among those obtained by the Levenshtein distance, the following Equation (6) is used, which introduces the left context of the word w_{err} in this distance. The most relevant solution w_i of the misspelled word w_{err} is given by:

$$w_i = \underset{v_k \in V}{Arg\min}\left(D_L(w_{err}, v_k) \Big/ \max_{i-n+1 \le l \le i-1} \Pr(v_k / w_l) \right) \qquad (6)$$

The importance of this approach consists in considering a history of size *n-1* and reducing the number of parameters to be estimated compared to the n-gram model.

6.3. Introduction of the N-Distant-Avg Language Model in the Levenshtein Algorithm

The following Equation (7) introduces the left context of the word w_{err} in Levenshtein distance, in order to find the right suggestion among those obtained by this distance. The most relevant solution w_i of the misspelled word w_{err} is given by:

$$w_i = Arg \min_{v_k \in V}\left((n-1)D_L(w_{err}, v_k) \Big/ \sum_{l=i-n+1}^{i-1} \Pr(v_k / w_l) \right) \quad (7)$$

Compared to the n-distant-max model, this model takes into account the case where the candidate word can be related to several words in the misspelled word's history.

7. Implementation

To test the efficiency of our two models, we used Kalimat as a training corpus and Python to code programs.

7.1. Introducing Used Data and Preprocessing

Kalimat is an Arabic multipurpose corpus that consists of six topics and contains 20, 291 articles. Table 1 shows the number of articles and words by topic:

Table 1. Kalimat corpus details.

Topic	Number of articles	Number of words
Culture	2782	1,359,210
Economy	3468	3,122,565
International news	2035	855,945
Local news	3596	1,460,462
Religion	3860	1,555,635
Sports	4550	9,813,366
Total	20,291	18,167,183

We have proceeded to the cleaning of this corpus from all punctuation marks. Then, we have extracted the vocabulary of the corpus, as well as the n-distant bigram transition dictionaries for different window sizes n in {2, 3, 4, 5, 6, 7, 8, 9, 10}. To note that n=2 is the particular case of bigram model.

A part of the corpus was used to create erroneous words to test our models. To do this, we have randomly chosen the target words to create errors of different types, taking into account all the editing operations (insertion, deletion, substitution and transposition) in generating errors. We have targeted the words which are not very short (word length > 2) and we have considered a limit of 3 as a maximum number of operations to be performed on the target word, to remain within the framework of spelling correction. All error creation operations are randomly chosen with a specification of a single editing operation for words of 3 characters, and 2 or 3 operations for words of 4 characters and more. Our test corpus includes 24,710 words of which 3,354 are misspelled.

7.2. Tests and Results

This part's purpose is to present the obtained results of the two proposed models and compare them to those of bigram model. To start with, we will first clarify how the size values relative to each model have been chosen.

7.2.1. Choice of the Window Size of the Two Proposed Models

To identify the optimal value of the size n of the two models n-distant-max and n-distant-avg, we calculated the correction accuracy for different values of n, and different values of m (number of the first suggestions returned). Tables 2 and 3 show these details:

Table 2. Accuracy rate (%) of n-distant-max models.

m	Bigram	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
1	26.63	64.61	70.22	73.64	74.72	75.88	72.78	70.19	66.58
1-5	55.75	89.21	92.16	92.84	94.40	94.51	93.83	92.99	91.26
1-10	67.44	94.22	95.77	95.95	97.50	97.67	97.44	96.63	95.65
1-15	73.44	96.33	97.73	97.47	98.63	98.81	98.63	98.15	97.38
1-20	77.25	97.14	98.51	98.36	99.26	99.37	99.20	98.99	98.42

Table 3. Accuracy rate (%) of n-distant-avg models.

m	Bigram	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
1	26.63	67.08	76.12	77.64	78.92	78.89	75.79	72.00	67.44
1-5	55.75	90.88	94.63	94.81	95.77	95.59	94.63	93.59	91.65
1-10	67.44	95.11	97.79	97.82	98.57	98.42	98.00	97.08	96.01
1-15	73.44	96.90	98.93	98.93	99.58	99.37	99.20	98.66	97.97
1-20	77.25	97.65	99.37	99.31	99.85	100.00	99.79	99.34	98.60

It is remarkable that the accuracy increases proportionally by incrementing the size n of the two models n-distant-max and n-distant-avg for any value of m from n=2 to n=6 or 7. Reaching the two size values 6 and 7, the accuracy starts decreasing for all the values of m. This shows very clearly that the optimum size value of n for the n-distant-max model is n=7 and for the n-distant-avg model is n=6. As a result, in the next part, we will keep only these two values.

7.2.2. Results and Comparative Analysis

- N-Distant-Max Model

The accuracy rate of 7-distant-max model, that of bigram model and that of edit distance are displayed in Table 4:

Table 4. Accuracy rate (%) of 7-distant-max, bigram and edit distance.

m	Edit distance	Bigram	7-distant-max
1	14.73	26.63	75.88
1-3	27.43	47.20	90.73
1-6	37.90	59.06	95.26
1-9	44.57	65.59	97.23
1-12	49.91	70.10	98.12
1-15	52.71	73.44	98.81
1-18	55.73	76.03	99.20
1-20	57.36	77.25	99.37
Total	57.36	77.25	99.37

From Table 4, we can observe that, in first position (m=1), the accuracy of the 7-distant-max model is much higher than that of the bigram model with a difference of 50% and much higher than the accuracy of the edit distance with a difference of 61%. If we take the first 12 solutions, the accuracy of the 7-distant-max model is always ahead with a difference of 48% compared to the edit distance and a difference of 28% compared to the bigram model. In total, the 7-distant-max model is better than the bigram model with a difference of 22% accuracy and better than the edit distance with a difference of 42%.

- N-Distant-Avg Model

Table 5 shows the accuracy rate of the 6-distant-avg model compared to 7-distant-max model, bigram model and edit distance.

Table 5. Accuracy rate (%) of 7-distant-max, 6-distant-avg, bigram and edit distance.

m	Edit distance	Bigram	7-distant-max	6-distant-avg
1	14.73	26.63	75.88	78.92
1-3	27.43	47.20	90.73	92.16
1-6	37.90	59.06	95.26	96.75
1-9	44.57	65.59	97.23	98.12
1-12	49.91	70.10	98.12	99.11
1-15	52.71	73.44	98.81	99.58
1-18	55.73	76.03	99.20	99.76
1-20	57.36	77.25	99.37	99.85
Total	57.36	77.25	99.37	99.85
Average time	**0.37 s**	**0.38 s**	**0.42 s**	**0.40 s**

From Table 5, it is seen that, in first position, the 6-distant-avg model shows a slightly higher accuracy than 7-distant-max model with a difference of 3%. This difference tends to 1% when the size of m increases. For example, it reaches 0.56% for the first 18 solutions. In total, we have a difference of 0.48% between the two models 6-distant-avg and 7-distant-max.

Regarding the average execution time per word per model, it increases slightly by increasing the size n. It is slightly higher for the two new models than edit distance and bigram model. This is explained by the increase of the number of entries in transition dictionaries in proportion to the size of window n.

The execution time of the 6-distant-avg model is slightly better than that of the 7-distant-max model with a difference of 0.02 s. this is explained by the complexity of the maximum function compared to the average function. It is also seen that from the thresholds, the n-distant-avg model has fewer parameters to estimate compared to the n-distant-max model.

8. Conclusions and Future Work

Throughout this article, we have managed to significantly improve the spelling correction of Arabic words in the light of their context, using two new models, n-distant-avg and n-distant-max, based on stochastic language models combined with edit distance. Both models improved spelling correction by more than 50% for the first position and more than 22% for the first 20 suggestions compared to bigram model. Our approach has the advantage of presenting better correction efficiency with a simplification of the parameters to be estimated. For future work, we would like to enhance the suggested spelling correction approach by taking into account the history of all previously corrected errors. We will also apply and test our approach on errors due to poor Optical Character Recognition of texts in Arabic, in particular, errors due to the deletion of the space between words.

References

[1] Aho A. and Corasick M., “Efficient String Matching: An Aid to Bibliographic Search,” *Communications of the ACM*, vol. 18, no. 6, pp. 333-340, 1975.

[2] Brucq D. and El Youbi A., “Représentation De Chaînes De Caractères Par Des Chaînes Induites De Markov,” *Actes RFIA’96*, pp. 651-658, 1996.

[3] Damerau F., “A Technique for Computer Detection and Correction of Spelling Errors,” *Communications of the ACM*, vol. 7, no. 3, pp. 171-176, 1964.

[4] Farra N., Tomeh N., Rozovskaya A., and Habash N., “Generalized Character-Level Spelling Error Correction,” *in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Maryland, pp. 161-167, 2014.

[5] Fossati D. and Eugenio B., “A Mixed Trigrams Approach for Context Sensitive Spell Checking,” *in Proceedings of 8th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico, pp. 623-633, 2007.

[6] Golding A. and Roth D., “A Winnow-Based Approach to Context-Sensitive Spelling Correction,” *Machine Learning*, vol. 34, pp. 107-130, 1999.

[7] Gueddah H., Yousfi A., and Belkasmi M., “Introduction of the Weight Edition Errors in the Levenshtein Distance,” *International Journal of Advanced Research in Artificial Intelligence*, vol. 1, no. 5, pp. 30-32, 2012.

[8] Gueddah H., Yousfi A., and Belkasmi M., “The

Filtered Combination of The Weighted Edit Distance and The Jaro-Winkler Distance to Improve Spellchecking Arabic Texts," *in Proceedings of 12th ACS/IEEE International Conference on Computer Systems and Applications*, Marrakech, pp. 16, 2015.

[9] Hadni M., El Alaoui O., and Lachkar A., "Word Sense Disambiguation for Arabic Text Categorization," *The International Arab Journal of Information Technology*, vol. 13, no. 1A, pp. 215-222, 2016.

[10] Hamza B., Abdellah Y., Hicham G., and Mostafa B., "For an Independent Spell-Checking System from the Arabic Language Vocabulary," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 1, pp. 113-116, 2014.

[11] Jelinek F., "Continuous Speech Recognition By Statistical Methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532-556, 1976.

[12] Jones M. and Martin J., "Contextual Spelling Correction Using Latent Semantic Analysis," *in Proceedings of 5th Conference on Applied Natural Language Processing*, Washington, pp. 166-173, 1997.

[13] Levenshtein V., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710, 1966.

[14] Nejja M. and Yousfi A., "A Lightweight System for Correction of Arabic Derived Words," *in Proceedings of Mediterranean Conference on Information and Communication Technologies*, Saidia, pp. 131-138, 2015.

[15] Nejja M. and Yousfi A., "The Context in Automatic Spell Correction," *in Procedia Computer Science*, vol. 73, pp. 109-114, 2015.

[16] Nejja M. and Yousfi A, "Correction of The Arabic Derived Words Using Surface Patterns," *in Proceedings of 5th Workshop on Codes, Cryptography and Communication Systems*, El jadida, pp. 153-156, 2014.

[17] Ringlstetter C., Schulz K., Mihov S., and Louka K., "The Same is Not The Same-Postcorrection of Alphabet Confusion Errors in Mixed-Alphabet OCR Recognition," *in Proceedings of 8th International Conference on Document Analysis and Recognition*, South Korea, pp. 406-410, 2005.

[18] Salton G. and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill Inc, 1986.

[19] Sharma S. and Gupta S., "A Correction Model for Real-Word Errors," *in Procedia Computer Science*, vol. 70, pp. 99-106, 2015.

[20] Zamora E., Pollock J., and Zamora A., "The Use of Trigram Analysis for Spelling Error Detection," *Information Processing and Management*, vol. 17, no. 6, pp. 305-316, 1981.

Saida Laaroussi is currently PhD student in the ES-Lab, 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.

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.

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.

Mohamed Nejja received his PhD in Computer Science and Engineering from the ENSIAS at Mohamed V University in Rabat, Morocco, in 2019. His areas of research interests include Natural Language Processing, Machine Learning, Artificial Intelligence.

Hicham Geddah is currently Associate Professor of Computer Science in the Department of Computer Science, ENS, at Mohammed V University in Rabat. He holds a doctorate in Computer Science from the ENSIAS at Mohamed V University in Rabat. The scope of his research covers: Natural Language Processing, Data Mining, Machine Learning and Deep Learning.

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.