Semantic Similarity Analysis for Corpus Development and Paraphrase Detection in Arabic

Adnen Mahmoud[1,2], and Mounir Zrigui[1]
[1]University of Monastir, Research Laboratory in Algebra, Numbers Theory and Intelligent Systems RLANTIS, Tunisia
[2]University of Sousse, Higher Institute of Computer Science and Communication Techniques ISITCom, Tunisia

Abstract: *Paraphrase detection allows determining how original and suspect documents convey the same meaning. It has attracted attention from researchers in many Natural Language Processing (NLP) tasks such as plagiarism detection, question answering, information retrieval, etc., Traditional methods (e.g., Term Frequency-Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), and Latent Semantic Analysis (LSA)) cannot capture efficiently hidden semantic relations when sentences may not contain any common words or the co-occurrence of words is rarely present. Therefore, we proposed a deep learning model based on Global Word embedding (GloVe) and Recurrent Convolutional Neural Network (RCNN). It was efficient for capturing more contextual dependencies between words vectors with precise semantic meanings. Seeing the lack of resources in Arabic language publicly available, we developed a paraphrased corpus automatically. It preserved syntactic and semantic structures of Arabic sentences using word2vec model and Part-Of-Speech (POS) annotation. Overall experiments shown that our proposed model outperformed the state-of-the-art methods in terms of precision and recall.*

Keywords: *Arabic language processing, word2vec, part-of-speech annotation, paraphrasing, semantic analysis, recurrent convolutional neural networks.*

Received January 24, 2019; accepted February 5, 2020
https://doi.org/10.34028/iajit/18/1/1

1. Introduction

The large amount of textual data shared on the web has facilitated the act of paraphrase. Its detection consists of determining a pair of sentences conveys the same meaning or not, by applying a semantic text similarity method [18]. Due to the complex problem of linguistic variations processing, this task has gained a significant interest in different Natural Language Processing (NLP) tasks such as plagiarism detection, question answering, document summarization, etc., [14]. Another issue is the lack of publicly available paraphrased resources in Arabic language.

To address these problems, we propose a paraphrased corpus based on both syntactic and semantic specificities of Arabic sentences. Because of the huge success of deep neural networks on features engineering, we employ Recurrent Convolutional Neural Network (RCNN) thereafter to detect Arabic paraphrases.

This paper is organized as follows: First, we present the state of the art, in section 2. Next, we describe the phases constituting the proposed methodology, in section 3. Subsequently, we detail the experiments carried out, in section 4. Finally, we give conclusions and future works, in section 5.

2. State of the Art

Corpora collection is a fundamental step to perform statistical analysis for different linguistic rules with minimal experimental-interference [8]. Researchers have been generated resources to investigate paraphrase detection problems according to two ways. The first one was simulate. It allowed obfuscating manually original sentences that their qualities were thereafter analyzed by experts. The second one was artificial. Different random obfuscation strategies were applied (e.g., word addition/deletion, word shuffling, synonym substitution, etc.,).

Microsoft Research Paraphrase Corpus (MRPC) was created using human annotations. It contained 5,800 pairs of English sentences in which 4,076 pairs were for the train and 1,725 pairs were for the test [19]. For example, Lee and Cheah [10] used MRPC dataset to conduct their experiments for English relatedness identification based on WordNet and cosine similarity techniques.

To evaluate the performance of plagiarism detection system, the PAN was an international competition providing a corpus including 20, 611 suspect and 20, 611 source documents. It contained rrandom obfuscation combinations such as paraphrase (i.e., synonyms, antonyms, and hyponyms substitution) and sentence reorganization. Several variants of this corpus

have been studied, let us quote: Daud *et al*. [5] used PAN-13 dataset composed by 3,653 suspect and 4,774 source documents. For English and Marathi language, Shenoy and Potey [23] proposed a fuzzy semantic similarity based on Naïve Bayes model. For experiments, they used PAN-PC-11 and PAN-PC-10 including 7,645 manual and 34,310 automatic paraphrases and PAN-PC-09 containing 17,127 artificial cases.

Zubarev and Sochenkov [24] used 7 million Russian documents. Different forms of plagiarism were applied that were varied from copy and paste to heavily disguised plagiarism.

Contrariwise, Sharjeel *et al*. [21] proposed a paraphrased corpusin Urdu language consisting of 160 documents: 65 source and 75 suspect. Texts were manually paraphrased by replacing words with their appropriate synonyms or different grammatical classes. To enrich it, they constructed an Urdu News Text Reuse Corpus (COUNTER) [22] composed by 1,200 documents with different levels of paraphrase (e.g., wholly derived, partially derived and non-derived). For paraphrase detection, various similarity methods were studied such as content, structure and style.

By cons, little attention has been considered for Arabic paraphrase detection because of its processing complexity. Arabic is clumpy and rich of specificities: writing from right to left [11, 15], diacritics and stacked letters above and below the baseline [9, 17], inflectional [4], derivational [6], and ambiguous [12].

Ameer and Juzaiddin [3] built a Corpus of Contemporary Arabic (CCA) composed by 415 texts. They modified manually 200 Arabic documents. Likewise, Al-Smadi *et al*. [2] collected Arabic tweets from Al-Jazeera and Al-Arabiya. They constructed 1,702 pairs for training and 791 pairs for testing. In addition, 526 training data while 253 testing data were paraphrased, respectively. For paraphrase detection, they used lexical overlap features with word alignment and topic modeling. Then, maximum entropy and support vector regression classifiers were employed. Furthermore, Nagoudi *et al*. [18] used an External Arabic Plagiarism Corpus (ExAra-2015). It was composed by different obfuscations forms like phrase and word shuffling, synonym substitution, diacritics insertion and paraphrasing.

Despite the richness of Arabic in terms of words construction and diversity meanings, there is no free Arabic paraphrased benchmarks. This has made most researchers to collect their own datasets from online sites. Hence, the need of Arabic paraphrased corpus development.

3. Proposed Approaches

3.1. Paraphrased Corpus Development

Figure 1 represents the proposed construction process of an Arabic paraphrased corpus:

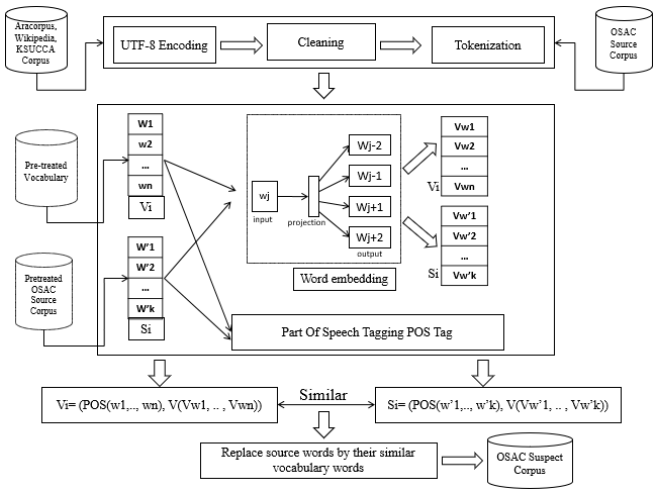

Figure 1. Proposed architecture of Arabic paraphrased corpus development.

- *Arabic documents collection*: A vocabulary model is created. It contains more than 2.3 billion words from the following Arabic corpora: Arabic Corpora resource (AraCorpus), King Saud University Corpus of Classical Arabic (KSUCCA) and Wikipedia. Furthermore, the Open Source Arabic Corpora (OSAC) is used as a source corpus from which passages of texts are extracted and replaced semantically from the vocabulary [20]. It contains 18,183,511 Arabic documents collected from different topics such as economics, history, sport, etc.
- *Documents preprocessing*: NLP techniques are useful to increase the performance and reduce the complexity required for Arabic paraphrase detection [6]. To reduce the incomplete and noisy data [13], unnecessary data are removed such as diacritics, extra white spaces, titles numeration, duplicated letters, non-Arabic words. Then, some ambiguous letters are normalized (i.e., Hamza "ٱ" and Taa "ة" to Alif "ا" and Ha "ه"). After that, text are splitted into tokens by regarding the white spaces or punctuations between words.
- *Part-Of-Speech (POS) annotation*: The relationship with adjacent words in a sentence are extracted to attach a part of speech tag to each one like noun, adverbs, quantifiers, etc., In this way, the grammatical features of Arabic sentences are preserved in which each word is annotated by its POS_i.
- *Synonym extraction*: The synonyms (Syn_1,...,Syn_k) of each source word are extracted from the vocabulary. To do this automatically, word-embedding (word2vec) algorithm based on skip gram model is employed. It is efficient to alleviate data sparsity problem and useful to offer an expressive semantic representation of words with fixed dimensional vectors. Formally, the middle word context x_v is predicted according to the unique representation of words in a surrounding window

size w_s, as denoted in Equation (1) [14]:

$$\frac{1}{V}\sum_{v=1}^{V}\sum_{-w_s\le i\le w_s, i\neq 0} \text{Log } p(x_{v+i}|x_v) \quad (1)$$

- *Paraphrased corpus generation*: The aim is that original and paraphrased sentences should have the same syntactic structure with semantically identical words that convey the same meaning, by proceeding as follows:

First, a random variable *v* restricted between a finite interval [β,*y*] is used to define the degree of reuse *P* to apply in the source sentence, as denoted in Equation (2):

$$g(x) = \begin{cases} \frac{1}{\gamma-\beta}, \beta \le v \le \gamma \text{ and } (\gamma-\beta) \in [1.33, \dots, 2.22] \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

Note that Pis fixed between 47% and 75% in this study. The number of words to obfuscate *O* in a sentence of words *N* is computed as follows in Equation (3):

$$O = N \times P \quad (3)$$

Random shuffle function is thereafter applied to replace each original word *wi* by the most similar one *Max* (*Syn*$_i$) that has the same POS. Subsequently, a weight φ of 0.1 is added to the highest score of similarity as shown in Equation (4):

$$\varphi(w_i) = \text{Max}(\text{Syn}_i) + 0.1 \quad (4)$$

3.2. Arabic Paraphrase Detection

The effectiveness of classical and deep neural networks models are studied. The aim is to determine which is the best in capturing sentences semantics and computing similarity.

- *Classical techniques*: traditional semantic vector models have focused on representing the words frequencies. In this study, several models are analyzed. Term Frequency-Inverse Document Frequency (TF-IDF) determines the relevance of a word to a document. It is computed as the product of the term frequency in the document and the logarithm of the reciprocal of its frequency in the corpus. Latent Semantic Analysis (LSA) analyzes the relationships between the terms and a set of documents. It learns latent topics by extracting low-dimensional semantic structure using Singular Value Decomposition (SVD) technique to get a low rank of word-document co-occurrence matrix. Furthermore, Latent Dirichlet Allocation (LDA) allows to associate a context with a document. Since a document is made up of a set of topics, LDA allows to assign words to each of them. In fact, paraphrase detection needs to assume that closed words in meaning should occur in similar pieces of text. This is by applying a semantic text analysis. However, these methods cannot analyze efficiently the changes of Arabic sentences especially when they are different (i.e., words co-occurrences are low or zero).
- *Deep neural network based techniques*: in recent decades, deep-learning based methods have been useful for learning hierarchical features and predicting more contextual information from sentences. Based on these advantages, we propose an architecture for Arabic paraphrase detection consisting of the following components (Figure 2):

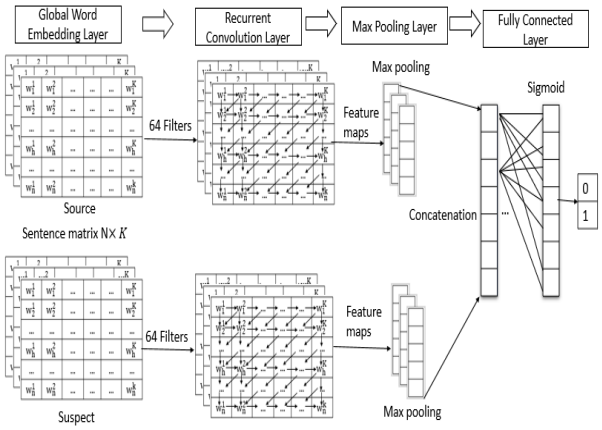

Figure 2. GloVe-RCNN model for Arabic paraphrase detection.

- *Global embedding*: the main advantage of word2vec algorithm is that it does better on analogy reasoning. However, the use of distant local contextual windows for its training presents a challenge with a large number of data. As a solution, we employ Global Vector Representation (GloVe) algorithm that combines the advantages of count based matrix factorization and contextual Skip gram model together. It exploits statistical information by training only on the non-zero elements in a word-by-word co-occurrence matrix. Then, an objective function *J* created a representative word vector with fixed dimensionality as denoted in Equation (5):

$$J=\sum_{i,j=1}^{N} g(X_{ij})\,(v_i^T \check{v}_j + b_i + \check{b}_j - \log(X_{ij}))^2 \quad (5)$$

Where: v_i and $\check{v}_j$ are the words vectors; b_i and $\check{b}_j$ are the scalar biases of the current word iand the context of the word *j*; *N* is the vocabulary and *g*(*x*) is the weighting function.

- *Convolutional Neural Network (CNN)*: GloVe embeds words of suspect and source documents into vectors that are used as entries in CNN. Given a sequence of words $w_{1:N}$, where each one is associated with an embedding vector w_i of dimension d For sentence modeling, a 1D convolution of width-kis the result of moving a sliding window-size over 2-4 word $sx_i = [w_i, \dots, w_{i+k-1}] \in R^{k\times d}$. Each convolution filter $C[i] \in R^l$ is a dot product between the concatenation of the embedding vectors in a given window and the weight vectors $U=[u_1,...,u_l] \in$

$R^{k.d\times l}$, with an addition of a bias term $b \in R^{l}$. Then, it is followed by a non-linear activation function Rectified Linear Units (ReLU) reducing the problem of data sparsity. It is defined as follows in Equation (6):

$$C[i] = \text{ReLU}(U.x_i + b) \quad (6)$$

Next, one max-pooling operation extracts the important key components from the sentence by taking the highest value observed in the resulting vectors, as denoted in Equation (7):

$$P_s = \text{Max}\{C[i]\} \quad (7)$$

Finally, a fully connected hidden layer learns complex nonlinear interactions and represents the global semantic information. To do this, the sigmoid function is the most suitable. It has a smooth gradient and generates values taken between 0 and 1. It helps CNN to learn efficiently at the back propagation process. It is defined as follows in Equation (8) [15]:

$$\text{output}(x) = \frac{1}{1+e^{-x}} \quad (8)$$

- *Recurrent Convolutional Neural Network (RCNN)*: CNN model based on a sliding window size is able to capture semantic meanings from the context of current words. Seeing that all the inputs and outputs are independent of each other, it fails in interpreting temporal information. Therefore, recurrent CNN RCNN structure came into existence to solve this issue. It is based on a hidden layer that remember previous information in the sequence. More precisely: a hidden state vector h_t is obtained by a nonlinear transformation with the previous hidden state h_{t-1} and the current input word x_t as denoted in Equation (9):

$$h_t = \text{ReLU}(U_{hh}h_{t-1} + W_{xh}x_t) \quad (9)$$

Where: U_{hh} is the weight matrix used to condition the previous hidden state h_{t-1}; x_t is the input vector at time t; W_{xh} is the weight matrix used to condition the current input x_t; and an activation function ReLU. Then, we employ the same max pooling and output fully connected layers that applied in CNN model.

4. Experiments and Discussion

4.1. Paraphrased Corpus Development

- *Word embedding parameters*: Figure 3 illustrates the performance evolution of the average (*Sen2vec*) of all words embedding's $\{w_1,...,w_n\}$ regarding different configurations of window sizes [2,…,7] and vector dimensions (250, 300, 350, 400, and 450), as defined in Equation (10):

$$\text{Sen2vec} = \sum_{i=1}^{n} \frac{w_i}{n} \quad (10)$$

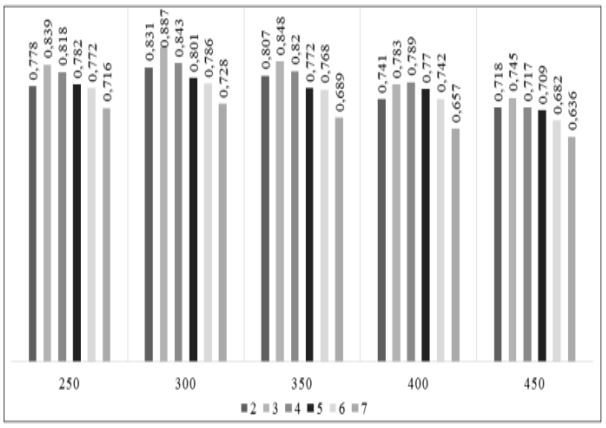

Figure 3. Sen2vec evolution regarding word2vec configurations.

For synonyms extraction, Table 1 presents the parameters of word2vec algorithm:

Table 1. Parameters of word2vec algorithm.

Parameters	Values
Vocabulary size	2.3 billion words
Vector dimension	300
Window size	3
Minimum count	≤ 5
Workers	8
Epochs	7

- *Example of paraphrased sentence generation*: The combination of word2vec and POS is efficient to preserve the features (i.e., syntactic and semantic) of Arabic sentence as shown in Table 2:

Table 2. Example of paraphrased sentence generation.

Source	لا تهمل عملك 'lā thml 'mlk' 'Do not neglect your work'		
w_i	**Similar words extraction**	**Word2vec**	**Word2vec-POS**
تهمل 'thml' 'neglect' 'verb'	Sim (w_i,اهتمام'āhtmām''care' 'noun') = 0.68	اهتمام 'āhtmām' 'care' 'noun' (0.73)	ترك 'trk' 'leave' 'verb' (0.67+0.1)
	Sim (w_i, ترك'trk''leave' 'verb') = 0.67		
	Sim(w_i, غادر'ġādr''quit''verb') = 0.73		
عملك ''mlk' 'work' 'complement'	Sim (w_i, وظيفة'wzyft''position' 'noun') = 0.70	وظيفة 'wzyft' 'position' 'noun' (0.70)	شغل 'šġl' 'job' (0.68+0.1)
	Sim (w_i, شغل 'šġl' 'job' 'complement') = 0.68		
	Sim (w_i, ktāb 'book'' object') =0.62		
Sen2vec		0.857	0.887

4.2. Paraphrase Detection Model

Global embedding parameters: The parameters of GloVe algorithm are presented in Table 3. For its training, different corpora (i.e., Akhbar Al Khaleej, AlWatan, King Abdulaziz City for Science and Technology (KACST) Arabic newspaper corpus, Arabic Gigaword, Kalimat, and International Corpus of Arabic ICA) are collected.

Table 3. Parameters of GloVe algorithm.

Parameters	Values
Size of co-occurrence matrix	1.119.436 × 1.119.436 words
Embedding size	300
Context size	3
Minimum occurrence	25
Learning rate	0.05
Batch size	512
Epochs	20

- *CNN parameters*: Table 4 presents RCNN parameters:

Table 4. Parameters of RCNN model.

CNN Layers	Parameters	
Convolution Layer	Filters Number	64
	Kernel Sizes	2, 3, 4
	Activation Function	ReLU
Pooling Layer	Type	Max-pooling
	Pooling Size	4
Fully Connected Layer	Activation Function	Sigmoid
	Loss	Binary cross entropy
	Optimizer	Adam
	Threshold	0.3

- *Performance metrics*: the evaluation of the proposed models is carried out using F-measure (F), as defined in Equation (11). It is the harmonic mean of the percentage of the relevant results (precision P) and the total percentage of correctly classified results (recall R) [16].

$$F = \frac{2 \times P \times R}{P + R} \quad (11)$$

- *Discussion*: the experimental results are summarized in Table 5:

Table 5. Experimental results.

Proposed Models		P%	R%	F %
Corpus	**Detection**			
Word2vec-POS	TF-IDF	69	72	70.46
	LDA	80.7	82	81,34
	LSA	75	77	75.98
	Word2vec	83.2	84.2	83.69
	GloVe-CNN	86	89	87.47
	GloVe-RCNN	88	91	89.47

- *Classical methods*: TF-IDF was slow for large vocabulary and was not able to extract the semantic relationships between words. In contrast, LDA reached the highest results compared to other classical techniques (i.e., TF-IDF and LSA) with 80.7% precision, 82% recall and 81.34% F-measure.
- *Neural networks*: although word2vec algorithm outperformed traditional methods and trained on distant local contextual windows, the statistics of the corpus could be misused. However, GloVe worked better on any number of data and represented efficiently the meanings of highly descriptive words.
- *Deep neural networks*: the use of GloVe and CNN together reached 87.47% F-measure. It comes down to the application of the window size in convolution layer. In contrast, it failed to interpret sequential information. Overall, experimental results demonstrated that GloVe-RCNN model achieved the highest results: 88% precision, 91% recall and 89.47% F-measure. It captured contextual data better than window-based structure in CNN. Seeing the complex nature of Arabic language, the quality of any paraphrase detection system depends on the adopted methodology. Table 6 and Figure 4 represent a comparison of experimental results with other existing approaches: Zubarev and Sochenkov [24] leveraged deep parsing techniques (e.g., TF-IDF, cosine, etc.,) for disguised plagiarism detection in Russian language. Daud *et al*. [5] detected copy and random obfuscations, by applying LDA and POS methods. Furthermore, Al-Anzi and Abyzeina [1] enhanced Arabic text classification using TF-IDF, LSA and cosine. Similarly, Nagoudi *et al*. [18] presented a method based on fingerprinting and word embedding. In addition, He *et al*. [7] used word2vec algorithm and CNN model for sentence similarity identification.

Table 6. Comparison study.

Systems	Methods	Corpora
Zubarev and Sochenkov [24]	TF-IDF + Cosine	SemEvalRus
Al-Anzi and Abyzeina [1]	LSA + TF-IDF + cosine	4,000 documents for train and 400 documents for test
Daud *et al*. [5]	LDA + POS	PAN-PC13
Nagoudi *et al*. [18]	Word2vec + word alignment	External Arabic Corpus
He *et al*. [7]	Word2vec + CNN	MSRP
Our model	GloVe + RCNN	OSAC source / suspect

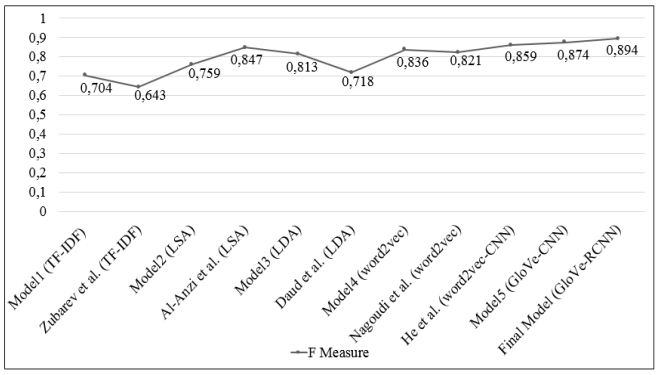

Figure 4. Comparison of experimental resultsregarding F1-measure.

5. Conclusions and Future Work

In this paper, we proposed a paraphrased corpus preserving both semantic and syntactic features of Arabic sentences. Original words are replaced by their synonyms having the same POS from a vocabulary. Different topologies of paraphrase are constructed (e.g., synonym substitution, add/deletion of words and

total copy). Experiments demonstrated that GloVe-RCNN model based on recurrent structure has achieved the highest results compared to the state-of-the-art methods. It was able to interpret sequential information and captured efficiently longer terms dependencies taking into account the structure of Arabic sentences. For future work, we would like to enhance the proposed Arabic paraphrase detection method and investigate the specificities of other deep learning architectures like Bidirectional RCNN, Long Short-Term Memory (LSTM), etc.

References

[1] Al-Anzi F. and AbuZeina D., "Toward an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing," *Journal of King Saud University, Computer and Information Sciences*, vol. 29, pp. 189-195, 2017.

[2] AL-Smadi M., Jaradat Z., AL-Ayyoub M., and Jararweh Y., "Paraphrase identification and Semantic Text Similarity Analysis in Arabic News Tweets Using Lexical, Syntactic, and Semantic Features," *Information Processing and Management*, vol. 53, no. 3, pp. 640-652, 2016.

[3] Ameer A. and Juzaiddin A., "Enhanced Tf-Idf Weighting Scheme for Plagiarism Detection Model for Arabic Language," *Australian Journal on Basic Application Sciences*, vol. 9, no. 23, pp. 90-96, 2015.

[4] Batita M. and Zrigui M., "Derivational Relations in Arabic Wordnet," *in Proceedings of 9th Global WordNet Conference*, Singapore, pp. 137-144, 2018.

[5] Daud A., Khan J., Nasir J., Abbasi R., Aljohani N., and Alowibdi J., "Latent Dirichlet Allocation and POS Tags Based Method for External Plagiarism Detection," *International Journal on Semantic Web and Information Systems*, vol. 14, no. 3, pp. 53-69, 2018.

[6] Haffar N., Hkiri E., and Zrigui M., "TimeML Annotation of Events and Temporal Expressions in Arabic Texts," *in Proceedings of International Conference on Computational Collective Intelligence, Auditorium Antoine D'Abbadie*, Hendaye, pp. 207-218, 2019.

[7] He H., Gimpel K., and Lin J., "Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Lisbon, pp. 1576-1586, 2015.

[8] Hkiri E., Mallat S., and Zrigui M., "Arabic-English Text Translation Leveraging Hybrid NER," *in Proceedings of 31st Pacifc Asia Conference on Language, Information and Computation*, Philippines, pp. 124-131, 2017.

[9] Hkiri E., Mallat S., Zrigui M., and Mars M., "Constructing a Lexicon of Arabic-English Named Entity Using SMT and Semantic Linked Data," *The International Arab Journal of Information Technology*, vol. 14, no. 16, pp. 820-825, 2017.

[10] Lee C. and Cheah Y., "Paraphrase Detection Using String Similarity with Synonyms," *in Proceedings of 4th Asian Conference on Information Systems*, Malisya, 2015.

[11] Mahmoud A. and Zrigui M., "Semantic Similarity Analysis for Paraphrase Identification in Arabic Texts," *in Proceedings of 31st Pacific Asia Conference on Language, Information and Computation*, Philippine, pp. 274-281, 2017.

[12] Mahmoud A., Zrigui A., and Zrigui M., "A Text Semantic Similarity Approach for Arabic Paraphrase Detection," *in Proceedings of 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, pp. 338-349, 2017.

[13] Mahmoud A. and Zrigui M., "Artificial Method for Building Monolingual Plagiarized Arabic Corpus," *Computacion y Sistemas*, vol. 22, no. 3, pp. 767-776, 2018.

[14] Mahmoud A. and Zrigui M., "Sentence Embedding and Convolutional Neural Network for Semantic Textual Similarity Detection in Arabic Language," *Arabian for Science and Engineering Journal*, vol. 44, no. 11, pp. 9263-9274, 2019.

[15] Mahmoud A. and Zrigui M., "Deep Neural Networks Models for Paraphrased Text Classifcation in The Arabic Language," *in Proceedings of the International Conference on Natural Language and Information Systems*, Salford, pp. 3-16, 2019.

[16] Mahmoud A. and Zrigui M., "Machine Learning Based Approach for Detecting Arabic Paraphrases," *in Proceedings of the International Business Information Management Association*, Granada, pp. 5035-5048, 2019.

[17] Mansouri S., Charhad M., and Zrigui M., "A Heuristic Approach to Detect and Localize Text in Arabic News Video," *Computacion y Sistemas*, vol. 23, no.1, pp. 75-82, 2018.

[18] Nagoudi E., Khorsi A., Cherroun H., and Schwab D., "A Two-Level Plagiarism Detection System for Arabic Documents," *Cybernetics and Information Technologies*, vol. 18, no. 1, pp. 1-18, 2018.

[19] Oussalah M. and Kostakos P., "On Web Based Sentence Similarity for Paraphrasing Detection," *in Proceedings of the 9th International Joint Conference on Knowledge Discovery, Engineering and Knowledge Management*, Funchal, pp. 289-292, 2017.

[20] Saad M. and Ashour W., "OSAC: Open Source Arabic Corpora," *in Proceedings of the 6th International Conference on Electrical and Computer Systems*, Lefke, pp. 1-6, 2010.

[21] Sharjeel M., Rayson P., and Nawab R., "UPPC - Urdu Paraphrase Plagiarism Corpus," *in Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, pp. 1832-1836, 2016.

[22] Sharjeel M., Nawab R., and Rayson P., "COUNTER: Corpus of Urdu News Text Reuse," *Language Resources and Evaluation*, vol. 51, pp. 777-803, 2017.

[23] Shenoy N. and Potey M., "Semantic Similarity Search Model for Obfuscated Plagiarism Detection In Marathi Language Using Fuzzy and Naïve Bayes Approaches," *IOSR Journal of Computer Engineering*, vol. 18, no. 3, pp. 83-88, 2016.

[24] Zubarev D. and Sochenkov I., "Paraphrased Plagiarism Detection Using Sentence Similarity," *in Proceedings of the International Conference Dialogue*, Moscow, pp. 1-10, 2017.

Adnen Mahmoud is a PhD student in the Higher Institute of Computer Science and Communication Techniques ISITCom, Hammam Sousse, Tunisia. He is member of Research Laboratory in Algebra, Numbers Theory and Intelligent Systems RLANTIS, Monastir, Tunisia. His areas of interest include natural language processing (Arabic language), machine learning, data mining and information retrieval. He has published many research papers in international journals and conferences.

Mounir Zrigui received his PhD from the Paul Sabatier University, Toulouse, France in 1987and his HDR from the Stendhal University, Grenoble, France in 2008. Since 1986, he is a Computer Sciences Assistant Professor in Brest University, France, and after in Faculty of Science of Monastir, Tunisia. He has started his research, focused on all aspects of automatic natural language processing (written and oral), in RIADI laboratory and after in LaTICE Laboratory. In addition, he is member of Research Laboratory in Algebra, Numbers Theory and Intelligent Systems RLANTIS, Monastir, Tunisia. He has run many research projects and published many research papers in reputed international journals/conferences.