

Artificial based Method for Building Monolingual Plagiarized Arabic Corpus

Adnen Mahmoud and Mounir Zrigui

LaTICE Laboratory, Department of Computer Science,
Unity of Monastir, Tunisia

Mahmoud.adnen@gmail.com,

mounir.zrigui@fsm.rnu.tn

Abstract. Plagiarism in textual documents is a widespread problem recently seen the large digital repository existing on the web. Moreover, it is difficult to make evaluation and comparison between solutions because of the lack of plagiarized resources in Arabic language publicly available. In this context, this paper describes our purpose of a paraphrased corpus construction automatically in order to deal with these issues and to conduct our experiments, as follows: First, we collected a large corpus containing more than 12 million sentences from different resources. Then, we cleaned it up unnecessary data by applying a set of preprocessing techniques. After that, we used word2vec algorithm to create a vocabulary from the collected corpus of which it was efficient to extract syntactic and semantic relationships between words to exploit. Subsequently, we replaced the words of source corpus with the vocabulary words most similar based on an index used randomly to eventually obtain a suspect corpus. Different experiments are done. Thus, we varied the dimensions of vectors and window sizes to predict the correct context of words and to identify the semantically closest words of the target.

Keywords: Arabic language, automatic creation, data collection, word embedding, paraphrase, plagiarism, semantic analysis.

1 Introduction

The reuse of ideas or words in whole or in part from someone else without granting the necessary credit of the source text copied. This phenomenon is of great interest to researchers in order to deal with it. However, Plagiarism detection in Arabic documents adds the difficulty of Natural Language Processing NLP. On the other hand, the lack of wide paraphrased corpora in Arabic language available publicly make hard the evaluation and comparison between proposed solutions. In this context, the fundamental thought of this paper is to build a monolingual Arabic plagiarized corpus

based on an artificial method. The rest of this paper is organized as follows: we begin by presenting the problem statement, in section 2. Then, we expose an overview of related corpora and their properties, in section 3. Thereafter, we describe our proposed plagiarized corpus construction, in section 4, and the experiments that we have done, in section 5. Finally, we end by a conclusion and our future works to achieve, in section 6.

2 Problem statement

The detection of plagiarism in textual documents is difficult because of the large amount of data exchanged in the web. This made unrecognized reuse much more prevalent. Thus, there are different levels of plagiarism, such as: word-to-word plagiarism, the obfuscation of the text content, and the reuse of the main idea independently of the words in the original text [1]. However, plagiarism in Arabic documents has attracted the attention of the research community seen the complexities of Arabic language that they make hard to detect plagiarism in them. It is a challenging language for many reasons: In addition, Arabic language is the language of the Koran and the sacred book of Muslims which is the 5th widely used language in the world [2, 3]. It is spoken by more than 422 million people as a first language and by 250 million as a second language [4]. Among the specificities of Arabic language which make its analysis more difficult, we mention:

- Words in Arabic language can mean a whole sentence resulting from an agglutination of the grammar. Thus, certain combinations of characters can be written in different ways which increases the ambiguity. It consists of a stem composed by a consonant root and a pattern morpheme, more affixes and enclitics. [5]
- The same text can be in several formats (fully vowel, semi-vowel, or unguarded). On the other hand, the lack of diacritic signs (vowels or diacritic points) in resources increases the problem of lexical and morphological ambiguity. Thus, the grammatical category of the ambiguous word can disambiguate its sense which complicates Arabic NLP. [6]
- A word may have several possible meanings and senses due to the richness of Arabic language: The possibility of changing words by their synonyms and reorganizing the structure of the sentence to another as active to passive and vice versa. [7]
- A word can has several lexical categories (noun, verb, adjective, etc.) in different contexts, which allows us to have different meanings of words. Consequently, this makes the Arabic script highly derivational and agglutinative. [8, 9]

To develop and evaluate a paraphrased plagiarism detection system, a large and increasing amount of digital texts is easily and readily available, making it simpler to

reuse but difficult to detect. However, we noticed that we need corpora in Arabic language with examples that initiate real plagiarism cases to make the evaluation and comparison of such solutions possible.

3 Related works

This section provides an overview of related works that deal with Arabic plagiarized corpora construction and their properties. Thus, plagiarized corpora mean the construction of a source corpus from which passages of texts are extracted and a suspect corpus in which the aforementioned passages are inserted after undergoing obfuscation processing. Therefore, we distinguish two ways to create corpora, distinguish: simulated corpora creation allows to authors or contributors to manually reuse another document based on different obfuscation types, such as: copy of words, or different word representations with and without diacritics in the case of Arabic language; and paraphrasing of which all kinds of modifications are applied, such as: restructuring, synonym substitution, etc., providing that the meaning of the original passage is maintained. However, artificial corpora creation allows the obfuscation strategies applied to fragments extracted automatically from source documents, such as: shuffling words/sentences, change the order of words, POS-preserving change of order, synonym substitution, and addition/deletion of words.

Several methods have been proposed to build corpora in different languages, we cite:

The volunteers have been encouraged to use their own knowledge of how to paraphrase a piece of text. Thus, in [1], it has been proposed a corpus in Urdu language containing in total 160 documents: 20 source documents have been original Wikipedia papers on well-known personalities and 140 suspicious ones have been manually paraphrased (plagiarized) versions produced by applying different rewriting techniques, such as: synonym replacement, changing intense or grammatical structure, summarizing content, splitting or combining sentence to make new ones. Also, text reuse has been occurred when one has borrowed the text either verbatim or paraphrased from an earlier written text. It has been the main idea in [10]: It has been shown the process of Urdu Short Text Reuse Corpus USTRC creation. This corpus has been contained 2,684 short Urdu text pairs, manually labeled as verbatim (496), paraphrased (1,329), and independently written (859).

In order to improve semantic text plagiarism detection, we distinguish also the PAN-PC corpus in English language which is a multi-dialect, expansive scale, open corpus of unoriginality, containing just artificial plagiarism occurrences. This corpus has been used in [11]. Thus, irregular appropriating has been emulated the initiatives of which a human would made to shroud duplicating, muddling through the reordering of the expressions, word substitution, equivalent word, and antonym utilize, erasures, and additions. The PAN-PC-10 corpus has been contained 27,073 text records, 15,925 arrangements of suspect reports and 11,148 arrangements of source archives produced

utilizing an artificial plagiarism program. Moreover, a single event may be reported in multiple articles in different ways that certain types of noun expressions such as names, dates, and numbers behave as anchors that may not change from article to another. In [12], a method for paraphrasing automatically from a corpus in Japanese has been described. Thus, expressions that conveyed the same information have been extracted using a simple co-referenced resolver to handle some additional anchors such as pronouns. Consequently, the resulting paraphrases have been generalized as models and stored for future use: 195 source documents and 100 paraphrased articles containing different cases of restriction: 106 co-reference or restriction, 37 co-reference and restriction and 32 co-references reference w restriction.

In contrast, we distinguished a semantic similarity search model for obfuscated plagiarism detection in Marathi language, in [13]. In this system, authors have been identified three datasets: the first two corpora PAN-PC-11 and PAN-PC-10 have been included 7645 manual paraphrases and 34,310 automatic paraphrases. On the other hand, PAN-PC-09 has been involved 17, 127 artificial cases but no simulated plagiarism cases have been found. For Marathi language, they have been collected some manual and artificial paraphrases for testing. Also, it has been described an approach to create a monolingual English corpus in PAN 2015 competition, in [14]. Two different obfuscation methods have been proposed to create different cases of plagiarism: The first has been an artificial method which it has been consisted of many obfuscation strategies, such as: synonym substitution, random change of order, POS preserving change of order and addition/deletion). The second has been a simulated method in which the SemEval dataset has been used for creating the cases of plagiarism by using pairs of sentences with their similarity scores.

However, we need to develop plagiarized corpora in Arabic language of which standardized evaluation resources are very beneficial to a wide range of field including information retrieval, NLP, etc. So, few works have been proposed: In [15], it has been built a plagiarized corpus for Arabic: The students have been asked to write an article about the importance of information technology and to use the Internet in their sources, especially in the case of a website. Different levels of plagiarism have been applied in the corpus, such as: exact copy, light modification or heavy modification. At the end, the proposed corpus has been composed by 1600 documents of which no plagiarized has been found. Likewise, the Corpus of Contemporary Arabic CCA has been used, in [16]. This corpus has been contained hundreds of documents in a variety topics and genres collected from magazines. It has been included in its corpus hundreds of source documents from Arabic Wikipedia. These documents have been collected manually by selecting documents that match the topics of the suspect documents. After, they have been incorporated in the corpus to baffle the detection, and only few cases have been created from them. After that, it has been realized that many of the collected Wikipedia articles (notably biographies) have been contained exact or near exact copies of large passages from the CCA documents, including: 1174 documents for the train and 1171 documents for the test.

4 Proposed artificial plagiarized Arabic corpus

In this section, we detail our approach for building a monolingual plagiarized Arabic corpus as shown in figure 1, including the following phases: First, we collected a large Arabic corpus from different resources. Then, we converted our dataset into text files (.txt) using UTF-8 encoding text. Thus, it is the standard form used for texts included in a corpora in order to facilitate subsequent processing's. After that, we applied different preprocessing methods to eliminate unnecessary data within text. Finally, we tried to create our suspect corpus automatically using word2vec algorithm and different random functions.

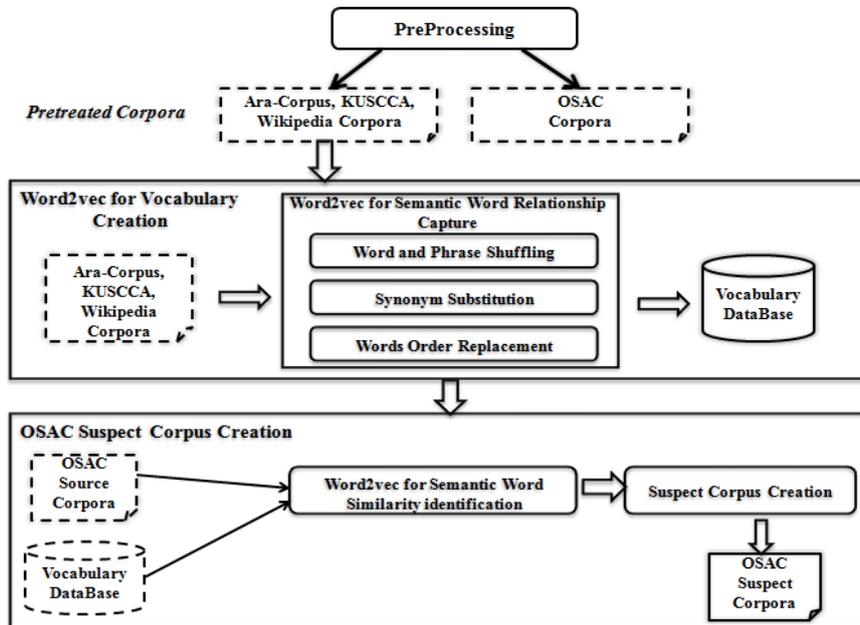


Fig. 1. Proposed Automatic plagiarized Arabic corpus

4.1 Preprocessing

We begin by cleaning our corpora used up unnecessary data by applying a set of pre-processing techniques in order to make subsequent processing's easier. Among the operations that we have applied, distinguish:

Corpora cleaning. Arabic is a rich and highly productive Semitic language, both derived and inflexive which complicated its analysis. So, we applied a set of operations to clean it by removing diacritics, extra white spaces, titles numerations, special characters, and no Arabic letters.

Corpora tokenization. The Arabic language is characterized by its cursivity and presents various diacritics. Thus, a word is composed by a set of related entities that are in turn formed of one or more characters. So, considering the complexity of Arabic language morphology and its analysis at the level of sentences and words, we apply the technique of tokenization to extract the words that can designate a whole sentence by separating the lemma and the word representing an assembly element of grammar, such as [17]: conjunctions, prepositions, definitional article or possession. Here is an example representing some tokenization forms in table 1:

Table 1: Tokenization forms examples

Forms	Arabic inflected form	English inflected form
Lemma + conjunction	ويخرج	<u>and</u> go out
Lemma + preposition	ليخرج	<u>to</u> get out
Lemma + definitional article	الخرج	<u>the</u> go out
Lemma + possession	خروجه	<u>his</u> go out

4.2 Train model

Corpora collection. We collected Arabic corpora from different resources to accomplish our work, characterized by different changes, such as: lexical, vocabulary and semantic, etc. Our collected corpus is composed by more than 2 billion words as shown in table 2. Among corpora that we have collected:

- Arabic Corpora Resource AraCorpus gives references of Arabic corpora computing linguistics of more than 126 million words. Among the corpora that we used: CNN, Thawra, Ahram, Akhbar, Almshaheer, Alquds, Alwatan, Al-watan, aps, Alsharq alawsat, and Attajdid.¹
- King Saud University Corpus of Classical Arabic KSUCCA represents a set of Arabic texts which are used in various types of computational linguistic research. It is composed by 50 million of words containing Arabic texts classified into 6 genres, such as: religion, linguistics, literature, science, sociology and biography. These genres cover the most of the topics that were popular in that period of time. [18]
- A set of Arabic papers from Wikipedia composed by more than 400,000 papers among the 295 active language editions, in February 2017.²

¹<http://aracorporus.e3rab.com/index.php?content=english>

²https://fr.wikipedia.org/wiki/Wikip%C3%A9dia_en_arabe

Table 2: Corpora collected

Source	Words number
AraCorpus	126.026.301
KSUCCA	48.743.953
Wikipedia	2. 158.904.163
Total	More than 2.3 billion words

Vocabulary creation. We create our vocabulary using word2vec algorithm in order to enrich the word representation according to its context (the words that surround it). Thus, words are projected to a continuous space of predefined size and words with similar contexts that can induce syntactic and semantic relations. In our work, we used the skip gram model to predict the possible context of words and to overcome the semantics ignorance problem of the data used compared to the Continuous Bag of Words CBOW model. Formally: given a sequence of words $\{w_1, w_2, w_3, \dots, w_T\}$, we try to maximize the average log probability:³

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

Where: k is the width of the context around each word w_t of which conditional probabilities are defined with Softmax function, as follows:

$$p(w_i | w_j) = \frac{\exp(u_{w_i}^T v_{w_j})}{\sum_{l=1}^V \exp(u_l^T v_{w_j})}$$

Where: V is the number of words in the vocabulary, u_{w_i} is the output representation of w_i and v_{w_j} is the input representation of w_j .

So, our main idea based on an analogy reasoning to obtain semantic relations between the words to exploit. Moreover, we try to determine the parameters that make our approach efficient by varying the dimensions of vectors and the window sizes by doing several experiments.

4.3 Test model

Source corpus. We used Open Source Arabic Corpora OSAC as a source corpus. It includes 22,429 text documents of which each text document belongs to 1 of 10 categories as shown in table 3, such as: economics, history, entertainments, education & family, religious and fatwas, sports, health, astronomy, low, stories, and cooking recipes. [4]

³<http://cedric.cnam.fr/vertigo/Cours/RCP216/coursFouilleTexte.html>

Table 3: Test corpus “OSAC” [4]

Document Category	Files Numbers
Economics	3102
History	3233
Education & Family	3602
Religious and Fatwas	3171
Sports	2419
Heath	2296
Astronomy	557
Low	944
Stories	726
Cooking Recipes	2373

Suspect corpus creation. In addition, we are talking about paraphrase when the contents of a text pair describe the same idea using different text rewrite operations, such as: addition or deletion of words (or sentences), synonym substitutions, lexical changes, active to passive switching, etc.

So, we proposed an artificial paraphrased corpus using the vocabulary that we have built based on word2vec algorithm, as follows:

1. First, we used random uniform function to identify the rate of plagiarism. After many experiments that we have done, we fixed the plagiarism rate P between 0.45 and 0.75 (Percent of plagiarism between 45% and 75%). Then, given the total number of words N in a given source document and the rate of plagiarism P to apply, the number of words to replace S is obtained as follows:

$$S = N \times P$$

2. Since an index of all the words of the original document varies between 0 and $N-1$, we try to find in which part of the document we will replace the words S according to a random index. Thus, we use a shuffle function to replace the order of sub-matrices, but their content remains the same.
3. After that, we try to create the suspect corpus, as follows: First, we extract all the words (tokens) from each sentence composed the source corpus. Then, we compare the words vectors representation of each token obtained with the words of our vocabulary model that we have built to extract their most similar words. In the case where a given token exists in our vocabulary model, we display a vector containing the most similar words for the current word, and we replace thereafter the original word by its most similar word to have thereafter a suspect corpus.

5 Experiments

To conduct our experiments, we used a large collection from different resources (Ara-corpora, KSUCCA, and Wikipedia papers) counting more than 2 billion words: First, we used them to build our vocabulary based on distributed word vectors representa-

tion. After that, we used the vocabulary that we have built as training model in order to exploit more concepts and to handle the rich of Arabic language morphology. Then, we evaluate the quality of our proposed methodology using OSAC corpus to generate the plagiarized corpus.

5.1 Train model

To perform the performance of our paraphrased corpus that we have created, we used an analogy reasoning using word2vec algorithm based on skip gram model. Thus, our main idea is to learn the semantic relationships between words to exploit according to their context in the level of the vocabulary and the suspect corpus creations. In this context, we experiment with different parameters, as follows: we have varied the parameters used in order to perform the suspect corpus creation using different vector dimensions, such as: 100, 250, 300, 350, and 500; and different window sizes to say the correct context of words, such as: [1, ..., 10]. Consequently, the best parameters that affect the resulting vectors to perform our experiments are: 300 as the vector dimension and 3 as the window size. Here is the following table 4 representing the final training configuration parameters that we have used to construct our proposed model:

Table 4 : Trainning configuration parameters

Parameters	Values
Vector Size	300
Min_Count	<= 5
Window size	3
Workers	8
Iterations Number	7

Where: vector size is the dimensionality of the word vectors; Min_Count is the words that have a total frequency less than 5 will be ignored; window size is the maximum distance between the current and predicted word within a sentence; workers is the train model using many worker threads; and iterations number is the number of iterations or epochs over the corpus.

5.2 Test model

Our main goal is how to build an efficient corpus to conduct our experiments for paraphrase detection. Thus, our proposed methodology allows replacing the words of OSAC source corpus with the vocabulary words most similar using an index randomly to eventually obtain OSAC suspect corpus. So, among the advantages of our proposed model, we mention: The use of Skip-gram model is an efficient method for learning a high quality of distributed vectors representation that capture a large number of precise syntactic and semantic relationships for better performance. Moreover,

the construction of a suspect corpus is done in an automatic way of which we distinguish different types of obfuscations, such as: total copy of the source document of which there is no obfuscation; shuffling of source words or sentences at random by changing their order randomly; replacing source words by their most similar words if they exist in the vocabulary.

Here is an example in table 5 illustrating our proposed automatic paraphrased corpus construction by dealing with the cases of creating a suspect sentence from a source sentence:

Table 5: Paraphrased sentence construction example

<i>Source sentence</i> : خرجت الجماهير المصرية في شوارع القاهرة فور انتهاء المباراة لتحتهفل بهذا الفوز : « The Egyptian fans went out in the streets of Cairo immediately after the match to celebrate this victory»	
Paraphrased sentence according to (vector dimension, window size)	
(250, 2)	
<i>Suspect sentence</i> : خرجت جماهير المصرية ضمن الشوارع الاسكندرية بعد بدء المباراة لتحتهفل بهذا فوز : “The Egyptian fans went out on the streets of Alexandria after the start of the game to celebrate this victory”	
(300, 3)	
<i>Suspect sentence</i> : خرجت جماهير المصرية في شوارع القاهرة عقب انقضاء المباراة لتحتهفل بهذا الانتصار : “Egyptian fans went out in the roads of Cairo after the match to celebrate the victory”	
(300, 4)	
<i>Suspect Sentence</i> : خرج جماهير المصرية ضمن شوارع القاهرة اثرانقضاء المباراة لتحتهفل بهذا فوز : “Egyptian fans took to the streets of Alexandria after the match to celebrate this victory”	
(350, 3)	
<i>Suspect sentence</i> : خرجت جماهير المصرية و الشوارع الاسكندرية قبل انتهاء المباراة لتحتهفل بهذا فوز : “Egyptian fans and the streets of Alexandria went out before the finishing of the game to celebrate this win”	

After several experiments of more than 1000 words according to different vector dimension and window sizes, we conclude that the parameters that help us to build the best paraphrased corpus were: 300 as a vector dimension and 3 as the window size. Here is an example in table 6 and figure 2 illustrating some paraphrased words according to the parameters that we have concluded and their most similar words found based on cosine similarity operation integrated in word2vec algorithm:

Table 6: Paraphrased words construction examples

Words	(vectors dimensions, window sizes)			
	(250, 2)	(300, 3)	(300, 4)	(350, 3)
"في: in"	ضمن: on 0.516	في: in 0.593	ضمن: to 0.542	و: and 0.532
“فور: immediately”	بعد: after 0.659	عقب: after 0.789	اثر: after 0.667	قبل: before 0.638
“انتهاء: after/finishing”	بدء: the start 0.651	انقضاء: the end 0.688	انقضاء: the end 0.577	انتهاء: the finishing 0.58

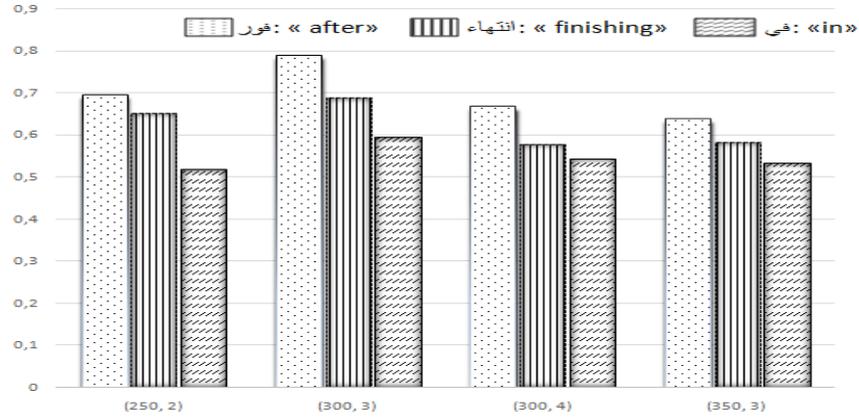


Fig.2. Examples of test configuration parameters

6 Conclusion

Paraphrase plagiarism is a significant problem and research showed that it is hard to detect it. Also, the lack of resources representing different cases of plagiarism especially in Arabic language complicates the evaluation of our proposed solutions. That's why; we proposed an artificial based method for building a plagiarized corpus to deal with it. Our main purpose is based on word2vec algorithm to create a vocabulary from the collected corpora of which it was efficient to extract semantic relations between words to exploit. Moreover, we used it to replace the words of source corpus with the vocabulary words most similar randomly to eventually obtain a suspect corpus. Experiments are shown that our proposed plagiarized corpus represent different cases of obfuscation, such as: total copy, source word/sentence shuffling, synonym substitution, paraphrase, etc. For future works, we try to improve the performance of our proposed method using Part Of Speech Tagging "POS Tag" technique. It allows providing morphological and syntactic annotations based on the grammatical class of word.

References

1. M. Sharjeel, P. Rayson, R. Muhammad, and A. Nawab: "UPPC - Urdu Paraphrase Plagiarism Corpus"; in proceedings of *Tenth International Conference on Language Resources and Evaluation LREC*, pp. 1832-1836, 2016.
2. A. Mahmoud, and M.Zrigui: "Semantic Similarity Analysis for Paraphrase Identification in Arabic Texts"; in proceedings of *the 31st Pacific Asia Conference on Language, Information and Computation, Philippine, PACLIC 31*, 2017.
3. S. Zrigui, A. Zouaghi, R. Ayadi, S. Zrigui and M. Zrigui, "ISAO: An Intelligent System of Opinion Analysis," *Research in Computing Science*, vol. 110, pp. 21-30, 2016.
4. M. K. Saad, and W. Ashour: "OSAC: Open Source Arabic Corpora"; in proceedings of *6th International Conference on Electrical and Computer Systems EECS'10*, Lefke, North Cyprus, 2010.

5. A. Boudhief, M. Maraoui, and M. Zrigui: "Elaboration of a model for an indexed base for teaching Arabic language to disabled people", *6th International Conference on CIST*, pp110-116, 2014.
6. O. Meddeb, M. Maraoui, and S. Aljawarneh: "Hybrid modeling of an offline arabic handwriting recognition system AHRS", *International Conference on Engineering & MIS ICEMIS*, Maroc, 2016.
7. S. Mallat, E. Hkiri, A. Zouaghi, M. Zrigui: Method of Lexical Enrichment in Information Retrieval System in Arabic. *International Journal of Information Retrieval Research(IJIRR)* vol. 3, no. 4, 2013.
8. A. Mahmoud, A. Zrigui, and M. Zrigui: "A Text Semantic Similarity Approach for Arabic Paraphrase Detection"; *International Conference on Computational Linguistics and Intelligent Text Processing CCLing*, 2017.
9. M. A. Ben Mohamed, S. Mallat, M. A. Nahdi and M. Zrigui, "Exploring the potential of schemes in building NLP tools for Arabic language," *International Arab Journal of Information Technology IAJIT*, vol. 6, no. 12, pp. 13-19, 2015.
10. S. Sameen, M. Sharjeel, R. M. A. Nawab, P. Rayson, and I. Muneer:" Measuring Short Text Reuse For The Urdu Language";*Language Resources & Evaluation*, vol. 51, issue 3, pp. 777-803, 2017.
11. A. H. Osman, O. M. Barukab:" SVM significant role selection method for improving semantic text plagiarism detection "; *International Journal of Advanced and Applied Sciences*, 4(8), pp. 112-122, 2017.
12. Y. Shinyama, S. Sekine:" Paraphrase Acquisition for Information Extraction", *Artificial Intelligence Review*, vol. 42, issue 4, pp. 851-894, December 2014.
13. N. Shenoy and M. A. Potey, "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.
14. S. Mohtaj, H. Asghari, V. Zarrabi:" Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus Notebook for PAN at CLEF 2015 ", 2015.
15. M. A. Siddiqui, I. H. Khan, K. M. Jambi, S. O. Elhaj, and A. Bagais: "Developing an Arabic plagiarism detection corpus"; *Computer Science & Information Technology (CS & IT)*, pp. 261–269, 2014.
16. I. Bensalem, I. Boukhalifa, P. Rosso, L. abouenour, K. Darwish, and S. Shikhi: "Overview of the AraPlagDet PAN@FIRE2015 Shared Task on Arabic Plagiarism Detection"; pp. 111-120, 2015.
17. S. B.Taamallah: "Prétraitement de données et création d'un segmenteur de l'arabe pour un système de traduction probabiliste vers le français";Université Stendhal Grenoble, Master 2012.
18. M. Alrabiah, A. Al-Salman, E. Atwell, and N. Alhelewh : "KSUCCA: A Key To Exploring Arabic Historical Linguistics" ; *International Journal of Computational Linguistics (IJCL)*, Vol. 5, Issue 2, pp. 27-36, 2014.