METHODS AND TOOLS FOR PLAGIARISM DETECTION IN ARABIC DOCUMENTS

Intisar Abakush

[1]Singidunum University,
32 Danijelova Street, Belgrade, Serbia

Abstract:
Due to the great revolution of data streams, people search for new ideas in miscellaneous fields of knowledge. As a consequence, there are various types of plagiarism issues. Excellent treatises have emerged with the aim of detection and protection from plagiarism. Statistics, methods, and software rwith more details and applications will prevent from overlapping and facilitate creating of something new in a clear and short manner. In our paper, we have focused on the plagiarism detection methods in Arabic documents and their systems. We have also highlighted some software which seemed to be useful in detecting plagiarized materials.

Key words:
plagiarism definition, plagiarism detection tasks,
plagiarism detection tools, Arabic.

1. INTRODUCTION

Due to the great extent of development in the world of technology and communication, plagiarism has become a significant challenge. Plagiarism has been found everywhere: on different levels of academic writing (school, institute, university, *etc.*), engineering, medicine, music, painting, literature, *etc.* It has been dubbed as illegal quotation, theft, cheating, and, piracy and alike.

2. PLAGIARISM AND ARABIC LANGUAGE

Plagiarism Definition

Derived from the Latin “plagiarius” which means “kidnapper, seducer, literary thief.”[1] From the earlier English word “plagiary” is “the one who takes someone’s words or ideas unjustly”.

For the time being, the word “plagiarism” does not have the unique term in Arabic. Current conditions are literary theft, scientific theft, arrogation, *etc.*, but there is a tendency of using the word “/انتحالintihal” which means arrogation of authorship.[2]

We may define plagiarism as an illegal quotation of someone else’s effort, whatever effort was it (an idea, invention, writing, methodology, design, *etc.*), and in different ways such as copy-paste function, by paraphrasing without exact citation.

Correspondence:
Intisar Abakush

e-mail:
ntsrbksh@gmail.com

DOI: 10.15308/Sinteza-2016-173-178

Arabic Language Characteristics

As aforementioned, the plagiarism problem is still a challenge, particularly owing to significant technological revolution. Still, it has been the biggest challenge in Arabic language.

The Arabic language belongs to Afro-Asian language group and it has been ranked as the fourth language in taxonomy of languages around the world. It has lots of specificities that make it so different compared to other Indo-European languages [3]. The Arabic language has many features and we have summarized them as follows:

- Arabic language has 28 characters [,أ, ب, ت, ث,..ي], three of them are vowels, [ا, ى, و] and others are consonants.
- A character's shape sometimes changes depending on its position in the word. For example: [ي، Y] in [See, رظني Yndr].
- Unlike other languages which are written from left to the right, Arabic language is written from right to the left and it does not have capitalization. There are two types of writing: ةعقرRogaa and خسنNasaha writing.
- Arabic documents are read and understood clearly by adding some diacritics above or under each character in word,[ً ُ ِ ْ ّ], for example [عَرَقKaraa, knock, while ,Kara عْرَق Zucchini]
- The root of every word in Arabic has just three characters, and new word is formed by adding some suffixes [name, verb, number, *etc*.]; for example: [Wrote, Ktabبتك], [بتكمmaktb. office].

Person and verb have three forms (singular, dual, and plural).

The remainder of this paper is organized as follows: the next section is related work, then plagiarism detection taxonomy; we classified the plagiarism into two definitions: plagiarism detection define types and plagiarism detection set tasks. Then, we demonstrated textual features in glance. After that Plagiarism Detection Tools in Arabic Documents, and plagiarism detection tools in Arabic documents, by highlighting some software as a good example of automatic plagiarism detecting in Arabic documents. The final sections presents concluding remarks.

3. RELATED WORK

The academic plagiarism is not a new phenomenon. Since year 1920, researchers have analysed the problem by focusing on North American colleges, where most of studies use self-report surveys to evaluate plagiarism behavior [4]. 1970 code clones and software misuse detection has started [5]. In 1990 plagiarism detection in natural languages was manual detection, while in 2005 researchers used automated plagiarism detection in text documents [5].

Since 2009, it actually started the work for detecting the plagiarism in Arabic world. We have to mention the biggest competition, by name "PAN plagiarism detection competition", where the evaluation corpora were mostly English. While "AraPlagDet"[Arabic Plagiarism Detection] share tasks is the first plagiarism detection competition on Arabic documents in 2015 and still continuous till now involves two sub-tasks, namely External and Intrinsic plagiarism detection[6].

"AraPlagDet"[Arabic Plagiarism Detection] is the first shared task that addresses the plagiarism detection in Arabic texts in "PAN plagiarism detection competition". Many researchers adopted this idea for their knowledge development and raising of the awareness level on the plagiarism problems and the importance of its detection in the Arab world.

4. PLAGIARISM DETECTION TAXONOMY

Plagiarism detection taxonomy sets two major types of plagiarism. First, one is: plagiarism detection sets types, and second one is: plagiarism detection sets approaches, which also called plagiarism detection tasks. Both also have two subtypes, as we shall soon find out.

Plagiarism detection set Types

As we can notice from the plagiarism definition and the related work, there are two types of plagiarism: plagiarism programming language and natural language plagiarism.

Plagiarism Programming Language the main attention of this type of plagiarism detection is tracking the metrics of that program, such as lines number, variables, data and part of program calls to another part of program in other program. There are many tools for detecting this kind of plagiarism [7]. For instance, but not limited:

- MOSS (Measure Of Software Similarity): It is free code plagiarism tool system for academic usage only. MOSS supports different operating system. The software uses finger print method to evaluate the similarity between evaluated codes.

- YAP/YAP3 (Yet Another Plague): A code-based it treats programs as a sequence of strings; the latest version YAP3 introduces an utterly novel algorithm for facing with the presence of block-moves in programs.
- SID (Software Integrity Diagnosis or Share Information Distance): Plagiarism detection system like MOSS and YAP proceeds with coding the input sequence and then comparing the coded sequences [8].

Natural Language Plagiarism deals with many textual features and diverse detection methods. Natural language plagiarism may also be called the textual plagiarism detection, and comprises two main classifications.

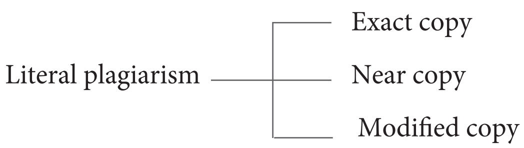
Plagiarism detection Language according to natural languages plagiarism detection, there are two types of lingual plagiarism detection, unilingual and multi-lingual plagiarism detection.

a) Unilingual plagiarism detection - most researches seek for developing the plagiarism detection system for unilingual plagiarism detection. It addresses the automatic identification and elicitation of plagiarism in unilingual, for example: Arabic-Arabic.

b) Multi-lingual plagiarism detection - researchers have been focused on this type of plagiarism detection just recently. It addresses the automatic identification and elicitation of plagiarism in multilingual contexts. For example, French - Arabic.

Textual plagiarism detection - is classified into Literal and Intelligent. Each of them has its sub-classification of plagiarizing and techniques of plagiarism detection.

Literal Plagiarism – the easiest and most common one, in which the plagiarist obviously copies the text from the original source and uses it as its won. Verbatim Plagiarism occurs in three cases: exact copy, near copy and modified copy (restructuring), with the last one being most challenging for detection [9].



Intelligent or masterly plagiarism is a grave fraud where the plagiarist tricks readers by presenting contributions of others as their own. Intelligent plagiarism appears in various intelligence phases such as manipulating text, translating text and adopting idea[10].

Text manipulation plagiarism - obscures the text manipulation mostly by changing its appearance, but not the idea. Words are being replaced by their synonyms/antonyms, and restructure the sentences in a text into shorter form.paraphrasing, by using a sentence reduction, etc. All of this is just one more form of plagiarism, unless being cited properly.

Translation is a form of plagiarism that occurs by original text translating from one language into another. This translation can be done automatically, by using some translating engine such as "Google Translator", or manually, by people who speak both languages.

Idea adoption is the most serious plagiarism that refers to the use of ideas of other people without citing the source of the idea. These could be results, contribution, findings, conclusions, *etc*.

Plagiarism Detection Tasks

Plagiarism detection is a hypernym for computer-based approach which supports identification, plagiarism detection information retrieval task supported by specialized IR (information retrieval) systems, called plagiarism detection systems which implement one of two generic detection tasks.

Extrinsic Plagiarism detection compares a suspicious document with a reference collection, which is a set of genuine documents. The comparison requires a document model with defined similarity criteria and the task is to retrieve all suspicious document [11].

Intrinsic Plagiarism Detection examine conditions of plagiarism by searching into doubting documents in isolation. Intrinsic plagiarism detection is highly percentage represented, human's ability to detect the plagiarism; by noting, analyzing different style of writing for the same author [12].

5. TEXTUAL CHARACTERISTICS

There are several textual characteristics to evaluate and characterize texts before applying a plagiarism method, especially quantifying according to plagiarism detection tasks and characterizing according to methods and tools used for detecting plagiarism in documents.

Textual characteristics in extrinsic plagiarism detection [13]

According to plagiarism detection tasks, textual features of representing documents in extrinsic plagiarism detection include:

Lexical characteristics, it works on character or word grams level. such as character n-gram and word n-gram, both called the fingerprints or shingles, in retrieval of text in detection of plagiarism in research.

Syntactic characteristics is plagiarism extraction by quantifying the similarity of sentences, phrases, part of speech, *etc.* the text in a syntactic way, such as conjunction of sentences, position of adverbs, preposition, and so on.

It is usually difficult to measure semantic similarity between documents, comparing with measuring just word similarity. And it useful when measured semantic similarity between documents to base on a similarity index that measures the number of similar words based on several possible algorithms [14]. These features and all previous are also called flat document features.

Structural characteristics also called tree features, reflected text formation, therefore detecting more documents semantics. We can find structural characteristic in header, title, sections, paraphrasing, *etc.* Structural characteristics could be used to create some web pages and special kinds of files, such as xml file.

Textual characteristics in internal plagiarism detection

According to Intrinsic Plagiarism detection tasks, textual features for representing documents in Intrinsic Plagiarism detection include just stylometric features. We know that Stylometry is extremely important in the context of internal plagiarism detection, and due to the truth that, each individual has its own specific writing style and hence, it is the only possibility to distinguish authors from each other [15]. Simon *et al.* defines Stylometry as "a discipline that determines authorship of literary works through the use of statistical analysis and machine learning" [15]. Textual characteristics fall into the following categories:

1. statistics of text: operate at the character level
2. Syntactic characteristic.: measure writing style at the sentence-level [15].
3. "POS characteristics: to quantify the use of word classes" [15].
4. "Closed-class word sets: to count special words" [15].
5. "Structural characteristics: which reflect text organization" [15]. According to these findings, each category refers to one specific text layer [15].

6. PLAGIARISM DETECTION IN ARABIC DOCUMENTS

Despite the lack of large-scale studies of the widespread plagiarism in the Arab world, this problem had attention from the large number of news which attest its pervasiveness. There are also some studies that show the lack of awareness on the definition and seriousness of plagiarism among Arab educative[16].

In the last years, many types of plagiarism detection research have been conducted, yet those concerning the text in Arabic language have remained quite limited. To the best of our knowledge, the sole works in this area are those of Alzahrani *et al.*, Menai *et al.*, and Jaoua *et al.* All of them used the external approach [17]. However, Intrinsic approach was the best reference of Bensalem, *et al.* As already mentioned, the greatest competition "PAN plagiarism detection competition" has widely opened the door to researchers for the methods development and plagiarism detection tools of the plagiarism in Arabic documents. "AraPlagDet" is the first common task that has been addressed to the plagiarism detection in Arabic texts.

These studies have suggested the use of plagiarism detection software as one of the problem solutions.

As we see in the table below, these are useful tools used to automatically detect the plagiarized Arabic documents, in good time and accurate way:

Turnitin/ Turnitout

This software is very good and its accuracy is high.

Turnitin has special and strict rules; if applied at university or faculty level, it achieves the best results. The disadvantage of this software is that it does not support individual work, respectively the system user has to be employed in some firm. Also, the user has to pay for every feature added in his system. Moreover, nowadays there is another software called Turnitout which works similar to Turnitin, , but it is intended for private users only. Although it is not deal with Arabic documents, turnitout gives good result in English materials.

Software	Language	Document extension
QARNET	Arabic, English	Microsoft Word, doc, txt, HTML, RTF
Turnitin/ Turnitout	31 language and Arabic, English	All files: power point, Excel, HTML, images, *etc.*
Ferret Copy Detection software	English, China, Arabic	Text documents (.txt) Word processor formats (.doc, .docx, .rtf, .abw) . and pdf documents (.pdf)
Aplag	Arabic	
Iplag	Arabic	
Plagiarism Checker	More than 190 languages supported.	Almost all files.

Table 1. plagiarism detection tools

PlagiarismChecker

Free, easy and detailed instructions, ideal for educators to check whether a student's paper has been copied from the Internet. The "Author" option allows a check if someone has plagiarized your work online.[18]

7. EVALUATION

We built our corpus *i.e.* suspicious documents adapted from various resources, like Bensalem, *et al.* Also, Almenai *et al.* Among these sources were web sites www.alwaraq.com (Al menai) and http: //ar.wikisource.org. (Bensalem) in order to evaluate automatic plagiarism detection methods and their precision and speed in the Arabic language.

For the same documents, we conducted our tests on three types of software (plagiarism.net, plagiarism detector, and QARNET). As shown in the following table and figure below, the results are almost close, but the time required showing the result is very different. Note that both software (plagiarism.net and QARNET) are considered as the most accurate gauge of plagiarism detection in Arabic documents, although Plagiarism detector software is faster than the hand to show the result.

Time/sec	Plagiarism%	Doc./Size	Software
109	58	Doc1/ 323word	Plagiarism. net
57	71	Doc2/Word148	
4800	57	Doc1/13	QARNET
	20	Doc2/12	
10	55	Doc1/14057	Plagiarism Detector
6	90	Doc2/12741	

Table 2. Plagiarism detection tools evaluation

The table discloses tha every software has its won technique to calculate the size of examined documents, the percentage of plagiarism alert threshold and the information retrieval sources (candidate documents), which are compared with suspicious documents to get the results as in the figure below.

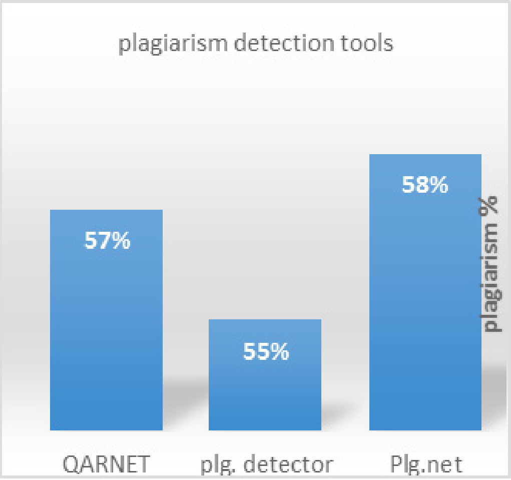


Figure 1. Plagiarism detection tools evaluation %

8. CONCLUSION

Methods that were developed and tested in Arabic documents are very few. As mentioned above, they were evaluated using different strategies and corpora, which makes them difficult to draw a clear conclusion on their performance. There was an effort to build annotated corpora in Arabic for external plagiarism detection and also Intrinsic plagiarism detection. So far, they have been used only by their authors [19]. In our study, we plan to improve our statistics of plagiarism detection in Arabic documents, to reach a new stable point at which the evaluation tools within the framework could be run smoothly from the box. In particular, we will encourage software submissions accurate and fast not only for detailed comparison but also for candidate retrieval, again using experimentation platform to facilitate this goal. Our vision is to implement accurate and fast automatic plagiarism detection evaluator, available to all researchers in this field.

REFERENCES

[1] Kashkur, M., Parshutin, S., & Borisov, A. (2010). Research into plagiarism cases and plagiarism detection methods. *Scientific journal of Riga Technical University. Computer science. Information technology and management science. Vol.*44. 139.

[2] Bensalem, I., Rosso, p., chichi, s. (2013). A new courpus of the evaluation of Arabic intrinsic plagiarism detection.CLEF2013, Valencia- Spain. Clef2013. clef-initiative.eu/ diapositive1-clef2013.pdf, p10.

[3] Menai, M. (2012*) Detection of plagiarism in Arabic documents.* (MECS). I.J. information Technology and computer science. Vol. 82. DOI: 10.5815/ ijitcs.2012.10.10

[4] Bela, G. (2014). *Citation based plagiarism detection. Springer fachmedien wiesbaden* 2014. Spriner vieweg. [dissertation otto-von-Guericke university, Germany, 2013

[5] Alzharni,S., Salim, N., & Abraham, A. (2012). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detect Methods. IEEE Transactions on systems. Vol.*42(2). DOI. 10.1109/ TSMCC.2011.2134847

[6] Bensalem. I., Boukhalfa, I., Rosso, P., Abouenour L., Darwish, K., A., & Chikhi, S. (2015). *Overview of the AraPlagDet PAN@Fire2015 shared task on Arabic plagiarism detection.* In fire2015 working notes papers. Gandhinger, India *Vol.* 114-117. DOI.

[7] Alzharni,S., Salim, N., & Abraham, A. (2012). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detect Methods. IEEE Transactions on systems. Vol.*42(2). DOI. 10.1109/ TSMCC.2011.2134847

[8] Kashkur, M., Parshutin, S., & Borisov, A. (2010). Research into plagiarism cases and plagiarism detection methods. *Scientific journal of Riga Technical University. Computer science. Information technology and management science. Vol.* 44. 139.

[9] Alzharni,S., Salim, N., & Abraham, A. (2012). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detect Methods. IEEE Transactions on systems. Vol.*42(2). DOI. 10.1109/ TSMCC.2011.2134847

[10] Alzharni,S., Salim, N., & Abraham, A. (2012). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detect Methods. IEEE Transactions on systems. Vol.*42(2). DOI. 10.1109/ TSMCC.2011.2134847

[11] Bela, G. (2014). *Citation based plagiarism detection. Springer fachmedien wiesbaden* 2014. Spriner vieweg. [dissertation otto-von-Guericke university, Germany, 2013

[12] Osman A., Salim N., and Abuobieda A. (2012). Survey of Text Plagiarism Detection. *Computer Engineering and Applications. Vol. 1, No. 1*

[13] Alzharni,S., Salim, N., & Abraham, A. (2012). *Understanding Plagiarism Linguistic Patterns, Textual Features, and Detect Methods. IEEE Transactions on systems. Vol.*42(2). DOI. 10.1109/ TSMCC.2011.2134847

[14] Alsmadi I., Alhami I., and Kazakzeh S., (2014). Issue Related to the Detection of Source Code Plagiarism in Students Assignments. *International journal of software Engineering and its Applications. Vol. 8. No.4. pp23-34.*

[15] Halvani O., Towards Intrinsic Plagiarism Detection. http://www. Halvani.de/math/pdf/(Oren_Halvani)_ Towards_ Intrinsic_ Plagiarism_ Detection.pdf.

[16] Bensalem. I., Boukhalfa, I., Rosso, P., Abouenour L., Darwish, K., A., & Chikhi, S. (2015). *Overview of the AraPlagDet PAN@Fire2015 shared task on Arabic plagiarism detection.* In fire2015 working notes papers. Gandhinger, India *Vol.* 114-117.

[17] Bensalem, I., Rosso, P., & Chikhi, S. (October. 2014), [pdf] Intrinsic plagiarism detection in Arabic text: preliminaryexperiments. *User.dsic.upv.es/~prosso/ resources/BensalemEtA1_CER12.pdf*

[18] Christopher, P, (2013, November). *Top 10 free plagiarism detection tools.* Retrieved October, 2015, from http://www.elearningidustry.com/top-10-free-plagiarism-detection-tools-for-teachers.

[19] Bensalem. I., Boukhalfa, I., Rosso, P., Abouenour L., Darwish, K., A., & Chikhi, S. (2015). *Overview of the AraPlagDet PAN@Fire2015 shared task on Arabic plagiarism detection.* In fire2015 working notes papers. Gandhinger, India *Vol.* 114-117.