Journal of Theoretical and Applied Information Technology
10th January 2016. Vol.83. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CROSS-LANGUAGE PLAGIARISM OF ARABIC-ENGLISH DOCUMENTS USING LINEAR LOGISTIC REGRESSION

[1]ZAID ALAA,[2]SABRINA TIUN, AND[3]MOHAMMEDHASAN ABDULAMEER

[1,2]Center for Artificial Intelligence Technology (CAIT)
Faculty of Information Science and Technology University Kebangsaan Malaysia, 43600 Bangi, Selangor Malaysia, and [3]Department of Computer Science, Faculty of Education for Women, University of Kufa, Iraq

E-mail:[1]zaidalaa92@yahoo.com, [2]sabrinatiun@gmail.com, [3]moh_ph135@yahoo.com

ABSTRACT

Cross-Language Plagiarism Detection (CLPD)is used to automatically identify and extract plagiarism among documents in different languages.The main challenge of cross-languageplagiarism detection is the difference of text languages, where the original source can be analysed and translated, and plagiarism can be detected automatically by comparing suspected text with the original text. This paper proposes an Arabic-English cross-language plagiarism detection method,to automatically detect the semantic relatedness between the words of two suspect targeted files.The proposed method consists of four phases. The first phase is a pre-processing phase,the second involves key phrase extraction and translation, the third phase used plagiarism detection techniques and the fourth phase is the classification process, which using Linear Logistic Regression (LLR). The evaluation process is created using precision and recall measurements of a dataset consisting of Wikipedia articles. The experimental resultsachieved96% precision, 85% recall and 90.16% F-measure. The results show that the LLRalgorithm can be used effectively to detect Arabic-English cross-language plagiarism.

Keywords: *Cross-Language Plagiarism Detection, Linear Logistic Regression,Arabic-English Cross-Language Plagiarism, Plagiarism Detection, And Wikipedia Articles.*

1. INTRODUCTION

Cross-language plagiarism detection attempts to automatically identify and extract plagiarism among documents in different languages. Plagiarized fragments can be translated. Verbatim copies may have their structuresaltered to hide the copying – this is known as paraphrasing and is far more difficult to detect. Online text publishing minimizes the difficulty of sharing and their reuse by other people. Some people copy text and reuse it without mentioning the authors. The huge amount of data that is provided by online internet resource networks maximizes the difficulty of detecting plagiarism effectively, as it requires more processing time. However, many types of data can be plagiarized, such as audio, text, images, and media clips[1].

The manual detection of plagiarism is inefficient for the large amounts of data that is published daily. Therefore, the automatic detection of plagiarism is necessary in order to protect the copyright of original author's work. However, plagiarism detection is not easy and requires a great deal of effort to detect, analyse, and report plagiarism efficiently using expert processes. Therefore, the automatic detection of plagiarism should be intelligent enough to handle the processes of detection accurately. For example, people can rewrite original texts in many styles to avoid plagiarism detection using manual or electronic methods i.e., 25 can be written as 'twenty five'[2].The study mainly focuses on the design and implementation of an Arabic-English cross-language plagiarism detection method, which automatically detects the semantic relatedness between the words of two suspected and targeted files. A Linear Logistic Regression (LLR) approach is proposedas a classification approach that is responsible for detecting plagiarism based on two binary possibilities. The rest of this paper is organized as follows: Section 2 describes work related to cross-language plagiarism detection; Section 3 shows the proposed technique;Section 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

explains the experimental results, and finally, Section 5clarifies the conclusion.

2. RELATED WORKON ARABIC-ENGLISH CLPD

This section provides an overview of related works that deal with the detection of cross-language plagiarism. Under this topic, Baroni and Bernardini[7].Conductedexperiments within a domain-specific corpus that consisted of English, Arabic, French, Spanish and Russian texts that were translated into Italian.

They employed the SVM classifier on lemmatized words and POS sequences and obtained the best accuracy through a combination of features including 1-gram word with tf-idf weighting, and 2-grams and 3-grams POS tags. They concluded that the task is dependent on the distribution of n-grams of function words and morpho-syntactic features.

In a related study, Pouliquen et al. [8] illustrated a statistical method that mapped multilingual documents into a language-independent document representation that gauged the similarity between mono and cross-lingual documents. Moreover, Anguita et al.

[9]introduced a cross-language plagiarism system for English-translated copies of Spanish document's detection. Their system was comprised of three stages; namely translation detection, internet search and report generation.

They classified text paragraphs with the help of supervised learning techniques (i.e., Support Vector Machines) as originally written in a specific language (N for Native language and F for Foreign language).

Furthermore, in a study conducted by Barron-Cedeno et al. [5] statistical methods were used to detect cross-lingual plagiarism. Specifically, they made use of the IBM Model 1 alignment model, fitted with a statistical bilingual dictionary, for the analysis of plagiarism in a parallel corpus. Initial studies in English and Spanish text fragments obtained satisfactory outcomes, but other experiments required a cross-lingual corpus for the evaluation phase.

In an extension of the work by Pinto et al. [10], English versus Spanish and English versus Italian documents were tested using the IBM Model 1 alignment model based on a bilingual statistical dictionary.

The system directly pinpointed the correlated words across different languages. The above studies indicate that alignment could be crucial to retrieval tasks involving cross-language information.

Also in the same line of study,was the work by Shiraz and Yaghmaee[11]. They introduced a method based on the overall dependence of textual contents that provided and employed the Vector Space Model (VSM). The method automatically detected bilingual plagiarism from English-Persian.In the context of Indonesian-English cross-language

plagiarism detection, Alfikri and Purwarianti[12] proposed a method consisting of three primary components, known as pre-processing, heuristic retrieval and detailed analysis. In a recent study, Omar et al. [13] demonstrated a plagiarism detection algorithm using both Arabic and English languages using the 'Bing' search engine.

The system supported both languages and used fingerprint and content comparison containing string-matching and tree-matching algorithms. The English publications obtained precision values of 80% while the Arabic publications obtained 90% precision.

Lastly, the pioneering Arabic-English cross-language plagiarism detection, using the Winnowing algorithm, was proposed by Aljohani et al. [3] to detect Arabic sentences translated from English sources without giving credence to the original authors.

3. THE PROPOSEDARABIC-ENGLISH PLAGIARISM DETECTION

The framework of the proposed Arabic-English plagiarism detection technique, which illustrates all stages of the process, can be seen in Figure 1.The proposed method consists of six major phases as follows:

i. Pre-processing
ii. Key phrases extraction and translation
iii. Retrieval of the candidate document
iv. Plagiarism detection techniques
v. Classification
vi. Evaluation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

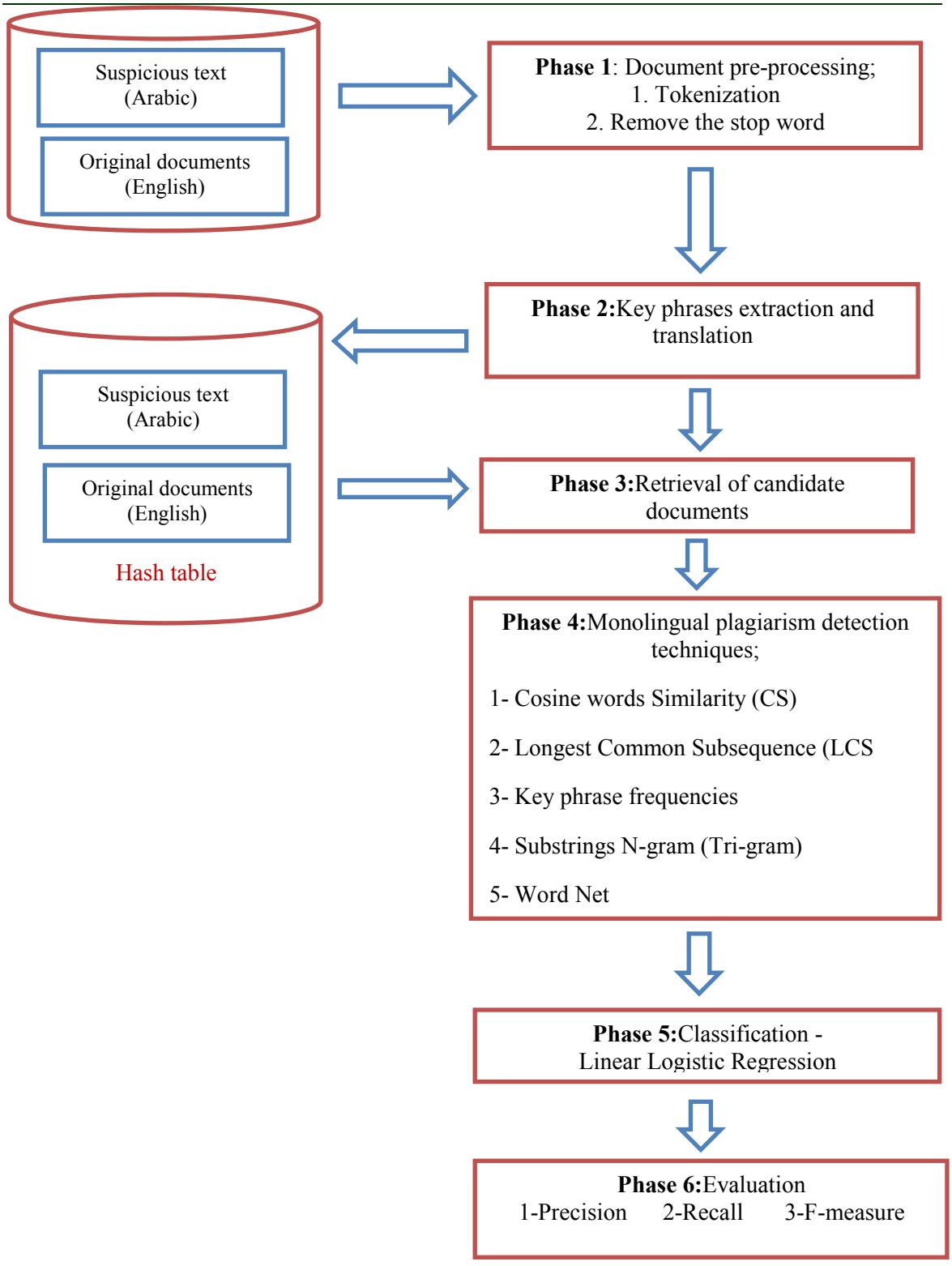

Figure 1: The proposed Arabic-English cross–languages plagiarism Detection method

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 Phase 1: Documents pre-processing

There are two main processes of the document pre-processing phase; (1) tokenization, (2) stop word removal (as shown in Figure 2). The main aim of this phase is to prepare the original document's dataset for similarity comparisons with other texts.

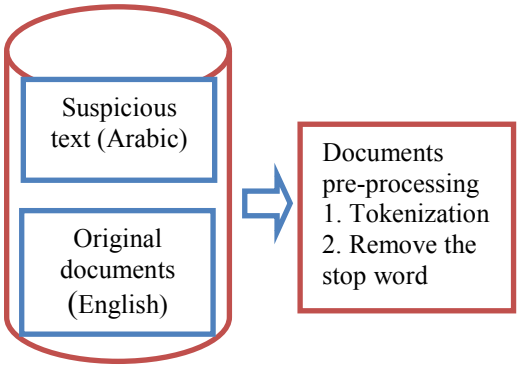

Figure 2: The pre-processing phase

A) TOKENIZATION

Tokenization is the process of breaking up a stream of text into words, phrases, symbols, or other meaningful elements; called tokens. The list of tokens becomes the input for further processing, such as parsing or text mining.

Tokenization is useful, both in linguistics (where it is a form of text segmentation) and in computer science (where it forms part of lexicalanalysis). Typically, tokenization occurs at the word level. However, it is sometimes difficult to define what is meant by a "word." Atokenizeroften relies on simple heuristics, for example:

i. Punctuation and whitespace may or may not be included in the resulting list of tokens.
ii. All contiguous strings of alphabetic characters are part of one token; likewise with numbers.
iii. Tokens are separated by whitespace characters, such as a space or line break, or by punctuation characters.

In languages that use inter-word spaces (such as most languages that use the Latin alphabet, and most programming languages), this approach is fairly straightforward.

However, even here there are many edge cases, such as contractions, hyphenated words, emoticons, and larger constructs, such as URIs (which for some purposes may count as single tokens).

In the proposed method, the tokenizerbreaks up a stream of input text of the Arabic file into words, phrases, symbols, or other tokens (as shown in Figure 3). Moreover, the English text resulting from the translation of the Arabic text is also broken up into tokens by using the same tokenization process.

(A) Arabic

Input	Output
اعلانقيام سلطة الشعب : هوالاعلان الذى اقره ممثلوالشعب الليبي تحت قيادةالعقيد [[معمر#القذافي]] في المؤتمرالقومي العام	"سلطة""قيام" "اعلان" "هو"""" :" "الشعب" الذى """ الاعلان" " "اقره" "الليبي""الشعب""ممثلوا " " قيادة""تحت""" "معمر """[[" "العقيد "القذافي ""#"" " "]] " في "القومي""المؤتمر""" "العام"

(B) English

Input	Output
The proclamation of the People's Authority: Is the declaration passed by the representatives of the Libyan people under the leadership of Colonel [Muammar Gaddafi]in General People's Congress	"The""proclamation ""of""the""People" "Authority""":""Is""" the""representatives ""of""the""Libyan" "people""under""th e"[Muammar Gaddafi] "leadership""of""C olonel""["""Muamm ar""Gaddafi""]""in ""General""People' s""Congress

Figure (3): Tokenization of the Arabic text (A) and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

English text (B).

Figure 3(a) shows how the tokenization has been done each word and component, including the stop words and special characters for the Arabic file. For example, the sentence "معمر القذافي"will become "معمر" and "القذافي". Meanwhile,Figure 3 (b) shows the same process for the English file,where "Muammar Gaddafi"

becomes "Muammar" and "Gaddafi." The space between words is used to separate the words and tokenize the paragraphs.

B) STOP WORDS REMOVAL

In natural language processing, stop words are words that are filtered out before or after the processing of natural language data (text). There is no single universal list of stop words used by all natural language processing tools; and indeed, not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase searches. Any group of words can be chosen as stop words for a given purpose. For plagiarism, some of the most common stop words are short function words, such as *the*, or in Arabic as *من, في، على*. In this case, stop words, if not filtered, can cause problems when searching for words that include them; particularly words such as 'من', 'في', or 'على' which mean "from", "in" or "on", respectively.

In addition, and regarding English text, after the tokenization step has divided the character sequence, the next step is to remove the stop words, which include preposition question and auxiliary verbs. Figure 5 shows the results of stop words removal. The list of the English stop words that has been used in this study is a default English stop words list, and is a well-known list used by many researchers, including [14]. In addition, the Arabic stop words list is taken from a study by [15].

(A) Arabic

Input	Output
"الشعب" "سلطة" "قيام" "اعلان" "الذى " " الاعلان" "هو" " :" " الليبي" "الشعب" "مثلوا" "اقره" "]]" " "العقيد " " قيادة" "تحت" في" "[]" " "القذافي " "#" "معمر " "العام" "القومي" "المؤتمر" "	"سلطة" "قيام" "اعلان " الاعلان" "الشعب" "ممثلوا " "اقره" " قيادة" " الليبي" "الشعب" " "معمر" "العقيد " "المؤتمر" "القذافي " "العام" "القومي"

(B) English

Input	Output
The proclamation of the People's Authority: Is the declaration passed by the representatives of the Libyan people under the leadership of Colonel [Muammar Gaddafi]inGeneral People's Congress	"The""proclamation""of""the""People" "Authority"":""Is""the""representatives""of""the""Libyan""people""under""the"[Muammar Gaddafi] "leadership""of""Colonel""[""Muammar""Gaddafi""]" "in""General""People's""Congress

Figure .4: Stop word removal of Arabic text (A) and English text (B)

After the tokenization has been done and the words, stop words and special characters have been identified, the stop words removal is applied. Figure 4 (A) shows the output of the stop words and special characters removal phase. The removed stop words are: في– تحت – الذي – هو, and the special characters are "]]" and "#". Meanwhile, Figure 4 (B) shows the output after removing the following stop words: "of," "the," "is," and "under" and the special characters "["and "]."

3.2 Phase 2 : Key phrase extraction and translation

The key phrase process is important for the coming stages where the extracted phrases will be stored in a hash table. The key phrase process is done for both files of Arabic and English documents. There are many challenges faced by the current methods

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of machine translation for the purposes of plagiarism detection, such as:

i. The high cost of translatingthe contents of all document.

ii. The difficulty of translating texts from one language into other languages.

iii. The machine translation methods cannot detect the lexical and structural changes of texts.

Therefore, the following steps are considered as efficient solutions forthe plagiarism detection of cross-languagesand have been followed in the proposed method.

Step 1: Analyse and classify documents based on the main key words of the document's topics automatically (as explained in Table 1 below). This step will divide paragraphs into words, and then prepare for the four sequences. The first sequence includes one word, which is the first word of the paragraph. The second sequence includes two words, which are the first word and the following word. The third sequence includes three words and the last sequence includes four words. The next step counts the frequency of the

first sequence, which includes one word only. Next, it looks for the frequency of the. second sequence, which includes two words from the paragraph. Then it looks for the frequency of the third sequence, which includes three words, and finally, it counts the frequency of the fourth sequence, which includes four words.

This step will be repeated many times by shifting the sequence to the second word and so on, and will be done recursively.

Step 2: Analyse and translate the key phrases of documents before comparing these phrases with other texts of the same translated language. In this step, each key phrase is translated and stored in ahash table. The translation is done using offline Google translator.

Step 3: Calculate the key phrase's frequencies in the selected texts. A high frequency is considered as an indication of plagiarism. The frequency is calculated by how many times the sequence repeats (as mentioned earlier).

Step4: The selected documents (that are suspect plagiarized) will be translated using machine translation to check the plagiarism efficiently.

Text documents can be characterized by a set of keywords giving an idea of what the text is about. The keyword definition can be quite complex; however, it is also generally intuitively obvious. Keywords or key phrases should briefly characterize what the actual text is about or refers to. So the keywords or the text key phrase of given text are quality words which refer to the actual text.

By extracting the phrases, we can get a list of key words from the source document. It is appropriate to ask where the key words come in the text. They can be globally spread or locally concentrated.The determination of the key phrases from Arabic text which would be checked is done using N-gram. Figure 6 shows the results of the key phrase phase.

An N-gram is a sub-sequence of n items from a given sequence. N-grams are used in various areas of statistical natural language processingandgenetic sequence analysis**.**

The items in question can be characters, words or base pairs according to the application. In this study, N refers to the number of words; where, in our case, four words are represented as the maximum sequence. For example, the sequence of words "proclamation" "People" "Authority"

"representatives" has a 4-gram of ("proclamation", "People", "Authority", "representatives"), and has a 3-gram of ("proclamation", "People", "Authority") an so on.For example, given an input sequence of w0 w1 w2 w3, in which this sequence refers to the words "proclamation," "People," Authority" and "representative," respectively. Table 1 depicts the example of how the given input sequence is extracted into key phrases using the n-gram technique.

Table 1: Example of key phrase extraction steps

Input Sequence	Sequence	Key phrase	Frequency
w0w1w2w3	w0		
….	w0w1	w0w1	3
	w0w1w2		
	w0w1w2w3		
	w1	w1w2	1
	w1w2		
	w1w2w3		
	w2		

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	w2w3 W3	w2w3	2

(A) Arabic

Input	Output
"سلطة""قيام" ""اعلان" " الاعلان" "الشعب" "ممثلوا " "اقره" " الليبي""الشعب" " "العقيد " " قيادة" "القذافي " "معمر" "القومي""المؤتمر" "العام"	سلطة الشعب معمرالقذافي إعلانقيامسلطةالشعب الشعباليبي العقيدمعمرالقذافي الإعلانالدستوري ثورةالفاتح الشعبالعربي الضباطالأحرار

(B) English

Input	Output
"proclamation" "People" "Authority" "representatives" "Libyan""people" "leadership" "Colonel" "Muammar""Gaddafi" "General""People's""Congress"	Peoples Authority Muammar Gaddafi proclamation Peoples Authority Libyan people Colonel Muammar Gaddafi Peoples Committees Constitutional Declaration Revolution Arab people Libya policy

Figure 5: Sample of key phrase extraction of Arabic text(A) and English text (B)

Phase 3: Retrieval of Candidate Documents

The main aim of this phase is to retrieve the document that matches the key phrase of the proposed plagiarism text. Therefore, the system does not need to retrieve all documents; thus saving processing time and cost. This technique uses the key phrase that was stored in the hash table.

Figure 6 shows the process of retrieving the candidate document only. It can be seen that storing the key phrase in the hash table plays a significant role in saving the time neededfor retrieving the whole document. The matching process is done by a function dedicated for this purpose in the proposed model and is based on the key phrases that have been extracted.

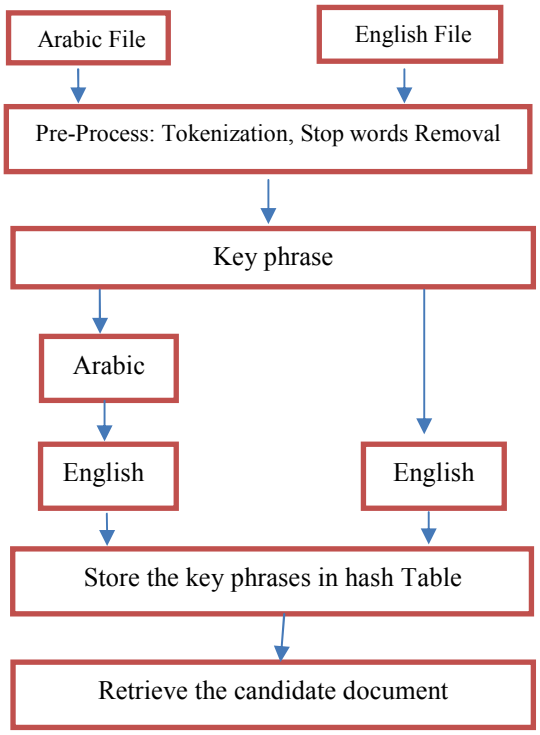

Figure 6: the process of retrieve the candidate documents.

A suspect document in Arabic goes through the key phrase extraction and the key phrase is translated into English after being extracted. It is then saved into the hash table. Retrieval is done according to the number of words (there is no need to retrieve the whole document and only the key phrase words are retrieved).

3.3 Plagiarism detection techniques

This phase measures the similarity between the original text and the suspected text. The features that can be analysed and measured in thisphase are Longest Common Subsequence (LCS), Cosine Similarity (CS), N-Gram and synonyms.

A combination of three algorithms, which are: Cosine Similarity (CS), Longest Common Subsequence (LCS), and N-gram, are then used to analyse the structure of sentencesto find similarities between short texts. Our similarity method is based on the semantic information of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

words in sentences to determine the degree of semantic equivalence between a pair of sentences. This involves three important stages (as shown in Figure 7).

First, word synonym is obtained from WordNet.

Second, we use synonyms for each word (word synonyms obtained from WordNet) in the reference document to generate all possible translation alternatives.

Third, the semantic equivalence is measured between two short documents by analysing the structure of sentences based on the three algorithms (CS, LCS, and N-gram).

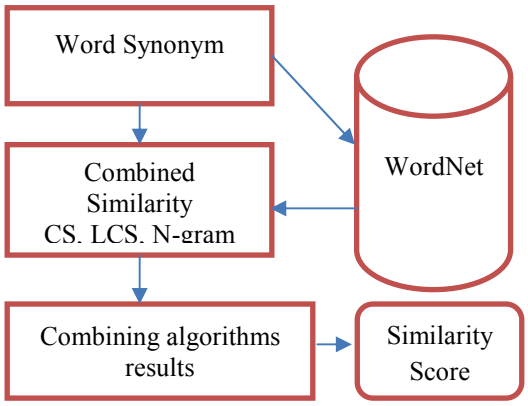

Figure 7: Plagiarism detection techniques phase

The overall similarity result of the method is calculated by combining the three methods to get the final similarity score between the test short document and the reference short document.

The score obtained from the combined three methods (CS, LCS, N-gram) should be between 0 and 1; where 0 means the short document is completely different and 1means the short document is identical.

The matching words synonyms phase includes the computation of semantic sentence similarity relying on the WordNet synonym, which takes all possible synonyms of each word's synonym that could be used by the translator in the test document, instead of the original words in the reference document, which shows that they both have same parts of speech and belong to the same set in WordNet, for getting the best score.

Similarity between short documents. In addition, WordNet can deal with synonyms as well and is a dedicated function of finding word synonyms.

In the 'measure similarity' phase, similarity measuring is carried out based on the applied similarity algorithms; Cosine words Similarity (CS), Longest Common Subsequence (LCS) and substrings N-gram (Tri-gram). All similarity measures are calculated as follows:

A) COSINE WORDS SIMILARITY (CS):

Equation 1 is the cosine similarity measure that is used to calculate the number of similar words that exist in sentence 1 (s1) and sentence (s2) to determine the score of similarity between them. Where $\sum_{j=1}^{m}(w_j w_j)$ is the number of similar words between the two sentences, and the total number of the weights of words in s1 and s2 is $\sqrt{\sum_{j=1}^{m} wj^2}\cdot\sqrt{\sum_{j=1}^{m} wj^2}$.

$$\frac{\sum_{j=1}^{m} w_j \, w_j}{\sqrt{\sum_{j=1}^{m} w_j^2 \cdot \sum_{j=1}^{m} w_j^2}} \quad (1)$$

In order to calculate cosine similarity between the two texts, they are transformed into vectors.

Each word in the texts defines a dimension in Euclidean space and the frequency of each word corresponds to the value in that dimension. The cosine similarity is measured by using the word vectors shown in Equation 1.

For example, a cosine similarity can be computed as below for two sentences; sentence1 and sentence2 as follows:

Sentence 1: Colonel Muammar Gaddafi
Sentence 2: Leader Muammar Gaddafi

Colonel 1 0
Muammar 1 1
Leader 0 1
Gaddafi 1 1

The two vectors are;

a: [1, 1, 0, 1]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

b: [0, 1, 1, 1]

$$CS\ (s1, s2) = \frac{1*0 + 1*1 + 0*1 + 1*1}{\sqrt{(1*1 + 1*1 + 0*0 + 1*1)^2}.\sqrt{(0*0+1*1+1*1+1*1)^2}} \approx 0.29$$

For example,for the sentences, 1: "Colonel Muammar Gaddafi" and 2: "Leader Muammar Gaddafi", after calculating the similarity among other sentences in the files and computing the result to be an input to the next step, which is LLR. Table 2 summarizes the whole calculation of Cs score between the two sentences.

Table 2: Results of CS similarity

Cosine words Similarity	S1:Colonel Muammar Gaddafi	S2:Leader Muammar Gaddafi
Value of CS (s1,s2)	0.29	

B) LONGEST COMMON SUBSEQUENCE (LCS):

Equation 2 is the longest common subsequence and compares the longest common substring between two character strings between sentence1 (s1) and sentence2 (s2).

$$\frac{2*|LCS\ (S_1,S2)|}{|S_1|+|S_2|} \quad (2)$$

For example, if Xis "MUAMMER" and Y is "MUAAMER". The longest common subsequence between XandY is "MJAU." Table 3 is generated by the function LCS, and shows the lengths of the longest common sub sequences between prefixes X andY. The *i*th row and *j*th column shows the length of the LCS between X_1andY_1...

Source String	proclamation" "People" "Authority" "representatives			
N-gram	1-gram: proclamation; People; Authority; representatives	2-gram: Proclamation People; People Authority; Authority representatives	3-gram: Proclamation People Authority; People Authority representatives	Proclamation People Authority representatives
No. of N-grams	1	2	3	4

Table 3: LCS sample values

		0	1	2	3	4	5	6	7
		Ø	M	U	A	M	M	E	R
0	Ø	0	0	0	0	0	0	0	0
1	M	0	0	0	0	0	0	1	1
2	U	0	1	1	1	1	1	1	1
3	A	0	1	1	2	2	2	2	2
4	M	0	1	1	2	2	2	2	2
5	M	0	1	1	2	3	3	3	3
6	E	0	1	1	2	3	3	3	4
7	R	0	1	2	2	3	3	3	4

The highlighted numbers show the path that the function backtrack would follow from the bottom right to the top left corner, when reading out an LCS. If the current symbols in X and Y are equal, they are part of the LCS, and we go both up and left (shown in bold). If not, we go up or left, depending on which cell has a higher number. This corresponds to either taking the LCS between X_{1..i-1} and Y_{1..j}, or X_{1..i} and Y_{1..j-1}.

C) SUBSTRINGS N-GRAM (TRI-GRAM)

(2)

Equation 3 is the N-gram that compares subs by substring to determine the number of similar substrings that exist in both sentence1 (s1) and sentence2 (s2). Where c is the number of common substrings between both sentences, and |s1| + |s2| is the total number of substrings in sentence1 (s1) and sentence2 (s2).

$$\frac{2*c}{|S_1|+|S_2|} \quad (3)$$

Let us consider a sentence consisting of the words; "proclamation" "People" "Authority" "representatives." The number of words available in this sentence is four. Therefore, one can extract N-gram of size maximum to four. The following table illustrates how many possible numbers of n-grams of the sentence.

Table4: The N-gram calculation process

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Hence, the maximum number of substrings of some specified size is the number of words in the sentence. Therefore, the number of words of the pointer table is sufficient for handling substrings.

D) THE COMBINATION OF SIMILARITY MEASURE

The first similarity measure, which is the Cosine similarity approach, focuses on words by using the term weights that are computed. The second similarity measure, which is the Longest Common Subsequence (LCS) algorithm, targets the longest common substring between two word strings, while the third approach, the N-gram algorithm, deals with the sequences of characters (character level Tri-gram) between sentences of two short documents (suspect and original documents). All three approaches use WordNet to get similarity information between word items.

Key phrases	LLR	Shared phrase	Share words
Libyan people	1		
Muammar Gaddafi	0	4	7
Revolution	1		
Arab people	1		
Free Officers	1		

In this paper, we propose to combine all three word similarity information approaches for use in our classification technique, the Linear Logistics Regression. The combination similarity word information is calculated by applying anequation that combines all three algorithms.

Com (s1, s2)= (Comcs (s1, s2) + Comlcs (s1, s2) + Com n-gram (s1,s2)) / 3 (4)

Moreover, LCS, N-Gram, CS results (which are numbers not words) are used as an input for LLR.

3.4 Phase 5: Classification - Linear Logistic Regression (LLR)

The classification processes can determine the plagiarism styles and levels of any text based on a set of features rules. This research adopts linear regression [17].[18] to predict plagiarism probabilities. The generated or extracted features from each document, such as word similarity, word synonyms, key phrase frequencies or N-gram, are helpful features in classifying a document as a plagiarism document or not.

The linear logistic regression is responsible for detecting plagiarism based on two binary possibilities; (1) plagiarized text, and (2) non-plagiarized text. The regression measures whether using two variables (shared words and shared phrase) will be chosen as the following equation (see Eq. 5):

$$logit(displag) = \beta_0 + \beta_1 x_1 + + \beta_k x_k \quad (5)$$

The linear logistic regression defines the predicted probability as in Eq. 6:

$$f(x) = P(displagirasized) = \frac{exp^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+exp^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} \quad (6)$$

where the coefficients β_i controls the effect of the predictor.The further a β_i falls from 0, the stronger the effect of the predictor x_i .Since our research problem is to detect certain parts of a document as being plagiarized or not, which is a binary classification problem; applying LLR in our work is therefore reasonable. In addition, with

good results obtained from previous works [3], to find the effect of LLR algorithm in plagiarism detection is very appealing.Table 5 shows the final results of the LLR process.

Table 5: LLR Results

3.5 Phase 6: Evaluation (Precision, Recall and - F-Measure)

A plagiarism detection system can be evaluated as a classification system; where each sentence belongs to one of two classes: plagiarized or original. In this study, three evaluation methods are used; precision, recall and F-Measure. The outcomes of plagiarism detection can be distributed as four types: true positive, true negative, false positive and false negative [16].True Positive (TP) is a set of plagiarized amounts already detected by thesystem. True Negative (TN) is a set of non-plagiarized parts and the system selects them as such. False Positive (FP) is a set of non-plagiarized parts, but the system detected them as plagiarized. False Negative (FN) is a set of plagiarized parts, but the system did not detect them. In terms of these four

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

sets, recall can be defined as follows: the recall measure is defined as the ratio of relevant plagiarized amounts detected by the system. Recall is a fraction of correctly categorized test cases divided by the number of test cases manually categorized as similar. The second performance metric is precision. The precision metric is used to measure the accuracy of the plagiarism detection system. The precision is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} (7)$$

4.EXPERIMENTAL RESULTS

The proposed model of this study was programmed with C# programming language. The objective of the proposed model is to solve the problem of plagiarism in Arabic text that may be copied from English text. Thus, several experiments are carried out to find the best setting (within our research scope and objectives) for Arabic-English CLPD system. In this research, a total of 318 Arabic files are used for both training and test. Arabic files are divided into paragraphs; then, going through the pre-process steps, translated into English. Next, extracting key phrases is done. Each file compared with the original of 54 English files(see Table 6).

Table 6: Detailed description of the experiment dataset

Dataset	Training	Test	Total
Arabic files	200	118	318
English files	54		54

As shown in Table 6, 200 Arabic files were used for training; where 118 were used as Test data. In addition, all English files were used for the comparison of both training and testing stages.

4.1 Experiment of combined Similarity measure and Evaluation

Presumably, having one document feature, a better accuracy of detecting plagiarism will be obtained due to the combining strength of all document features. In our case, by combining the mentioned features; CS, LCS and N-gram, we will be able to get better results than using individual features only.

Therefore, in Experiment IV, we will use combined similarity measures as features to be used in LLR.In order to evaluate the impact of the Precision (positive predictive) is the parts of retrieved documents that are relevant; whereas recall (sensitivity) is the parts of relevant documents that are retrieved from the corpus. A high precision value refers to effectiveness and efficiency, while a high recall value refers to durability [16]. The third metric is F-measure. F-measure combines precision and recall into a single measurement to balance them.The range of F-measure is between 0 and 1. A combination of both measures (recall and precision) offers a better picture of an obtained result [2].

combined features used in LLR classification, the value of a pair of extracted key phases of the combined feature is computed. A sample of such output, based on Eq. (4) above, is shown in Table 7. The value of the combined features should be in the range 0 to 1.In our experiment, the values of the combined features (far right column) ranged from 0.2 to 8. In Table 7, we also included the value ofthe individual feature, so that one can see the different range values.

Table 7: Sample of the individual feature values and the combined feature

Suspicious Text (Arabic)	Original documents (English)	Cosine Similarity	Longest Common Subsequence	N-gram (Tri-gram)	Combined
A (10).txt	E (10).txt	0.051	0.222	0.713	0.328
A (100).txt	E (100).txt	0.058	0.58	0.882	0.506
A (101).txt	E (101).txt	0.073	0.228	0.988	0.429
A (102).txt	E (102).txt	0.042	0.122	1	0.388
A (103).txt	E (103).txt	0.104	0.272	0.87	0.415
A (104).txt	E (104).txt	0.051	0.466	0.653	0.39
A (105).txt	E (105).txt	0.075	0.039	0.661	0.258

The performance of the combined features used in LLR is measured by Precision, Recall, and F-measures, and the results are shown in Table 9. The results presented in Table 7 are of an individual file. The 'shared phrase detected' mentioned in Table 8 refers to the plagiarism of key phrases detected by the LLR.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The previous values shown in Table 7 are the input of feature (combined features) into the LLR and the results are shown in Table 8.

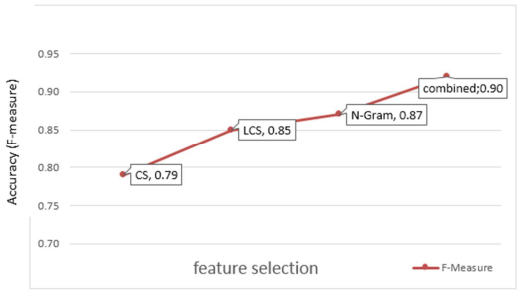

Figure 8: Evaluation comparison of CS, LCS, N-gram and combined.

Finally, the final evaluation values for the combined features for the whole dataset's similarity are shown in Table 9.

Table 9: Combined features evaluation final results

Precision	**0.96**
Recall	0.85
F-Measure	0.90

The implementation of the proposed model shows that different values (features) have gained different results. Table 10 shows the results of all methods (features) including the combined features.

Table 10: Evaluation results of the implemented features

Evaluation	**CS**	**LCS**	**N-Gram**	**Combined**
Precision	0.75	0.92	0.84	0.96
Recall	0.84	0.78	0.90	0.85
F-Measure	0.79	0.85	0.87	0.90

As shown in Table 10 and Figure 8, the combined features gained the highest value of F-measure (90%) compared to when individual features were used in the LLR.N-gram with F-measure of 87% was slightly higher than the LCS with only 85% of F-measure. The lowest accuracy of detecting plagiarism using LLR was CS with only 79%.

Obviously, the obtained results support our assumption that combining all features of documents will get a better result in detecting plagiarism than using single individual features. Therefore, the obtained results answer our second objective, which is to see the impact of selected features and the combination of the selection features, when used in Linear Logistic Regression (LLR) of Arabic-English plagiarism detection model. Furthermore, our obtained results were better than the previous work of [3] at only 89%

Table 8: Sample of combined feature individual file result Sample of combined feature individual file result

Case	Arabic	English	No. of plagiarized shared phrases detected manually	Shared phrases detected	Results				
					TP	FP	FN	Recall	Precision
1	A (10).txt	E (10).txt	5	4	4	0	1	0.80	1
2	A (100).txt	E (100).txt	28	24	23	1	4	0.85	0.95
3	A (101).txt	E (101).txt	7	6	6	0	1	0.85	1
					9	0	2	0.82	1
					20	1	3	0.87	0.95
					6	0	1	0.85	1
7	A (105).txt	E (105).txt	16	14	13	1	2	0.86	0.92
8	A (106).txt	E (106).txt	13	11	11	0	2	0.84	1
9	A (107).txt	E (107).txt	11	9	9	0	2	0.82	1
10	A (108).txt	E (108).txt	18	16	15	1	2	0.88	0.93
11	A (109).txt	E (109).txt	8	7	7	0	1	0.87	1
12	A (11).txt	E (11).txt	5	4	4	0	1	0.80	1
13	A (110).txt	E (110).txt	13	11	10	1	2	0.83	0.90
14	A (111).txt	E (111).txt	28	25	24	1	3	0.89	0.96
15	A (112).txt	E (112).txt	14	12	11	1	2	0.85	0.91

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(as shown in Table 11), and by the work of [13] with precision at 90% - which is 6% lower than the results obtained in this study (precision = 96%).

Table 11: Comparison of the gained results with previous works

Method	Precision	Recall	F-Measure
Proposed	0.96	0.85	0.90
Aljohaniet al. (2014)	0.97	0.81	0.89
Omar et al (2013)	90%	———	———

5. CONCLUSION

This paper has presented an automatic Arabic-English cross-language plagiarism detection method. The proposed method consists of four phases: the first the pre-processing phase, the second phase is key phrase extraction and translation, the third phase is plagiarism detection techniques and the fourth phase is the classification process using Linear Logistic Regression (LLR). The experiments of this study included the implementation of three combined algorithms, which are: Cosine Similarity (CS), Longest Common Subsequence (LCS), and N-gram. These were used to solve the problem of plagiarism. Each method was tested individually and the combination method gave the final result of the proposed system. In conclusion, the combination of these measurements gave the best performance. The evaluation of the proposed methods was done using three measures: Precision, Recall and F-Measure. The evaluation by the experimental results shows that the best performance measure was obtained using a combination of the three algorithms together. The results have provided a similarity measure that shows a significant correlation to human intuition.

6. ACKNOWLEDGEMENT

This researchprojectis fundedby Malaysia Government under research Grant ERGS/1/2013/ICT07/UKM/03/1.

REFERENCES

[1]Marc Franco-Salvador, Parth Gupta, Paolo Rosso. Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. Bridging Between Information Retrieval and Databases,Lecture Notes in Computer Science Volume 8173,(2014), pp.227-236.

[2] Barrón-Cedeño, A., Gupta, P., &Rosso, P.Methods for cross-language plagiarism detection. Knowledge-Based Systems, 50,(2013), PP. 211-217

[3] Aljohani, A., &Mohd, M. Arabic-English Cross-language Plagiarism Detection using Winnowing Algorithm. Information Technology Journal, 13(14), 2349.(2014). PP.1-10.

[4] Barrón-Cedeno, A., Rosso, P., Pinto, D., & Juan, A. On Cross-lingualPlagiarism Analysis using a Statistical Model. In PAN.(2008), July, PP. 1-10.

[5] Barrón-Cedeno, A., Rosso, P., Pinto, D., & Juan, A. On Cross-lingual Plagiarism Analysis using a Statistical Model. In PAN.(2008, July). PP.1-5.

[6]Chong, M. Y. M. A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. (2013). PP.1-10.

[7] Baroni, M. and Bernardini, S. A new approach to the study of translationese: Machinelearning the difference between original and translated text. Literary and Linguistic Computing, 21(3), (2006). PP.259–274.

[8] Pouliquen, B., Steinberger, R., &Ignat, C. Automatic identification of document translations in large multilingual document collections. arXiv preprint cs/0609060.(2006). PP.1-10.

[9] Anguita, A., Beghelli, A., &Creixell, W. Automatic Plagiarism Detection. In *NLPKE 2011,7th International Conference on Natural Language Processing and Knowledge Engineering*.(2011).PP.1-10.

[10] Pinto, D., Civera, J., Barrón-Cedeno, A., Juan, A., &Rosso, P. A statistical approach to cross lingual natural language tasks. Journal of Algorithms, 64(1), 51-60.(2009). PP.1-10.

[11]FarzinYaghmaee, SorayaEnayati Shiraz. Introducing an Automated Technique for Bilingual Plagiarism detection of English-Persian Documents, Current Trends in Technology and Science , 8thSASTech 2014 Symposium on Advances in Science & Technology-Commission-IV Mashhad, Iran,2014.pp.28-32.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[12]ZakiyFirdausAlfikri, AyuPurwarianti.The ConstructionofIndonesian-English Cross LanguagePlagiarism DetectionSystem UsingFingerprintingTechnique, Vol. 5, No 1,journalofcomputerscienceandinformation ISSN:2012.PP. 2088-7051.

[13] Omar, K., Alkhatib, B., &Dashash, M. The Implementation of Plagiarism Detection System in Health Sciences Publications in Arabic and English Languages. International Review on Computers & Software, 8(4).(2013).PP.5.

[14] Tong, Simon, Uri Lerner, Amit Singhal, Paul Haahr, and Steven Baker. "Locating meaningful stopwords or stop-phrases in keyword-based retrieval systems." U.S. Patent 8, issued July 3, 2012. PP.214-385.

[15] Mahdaouy, A. E., Ouatik, S. E., &Gaussier, E. A Study of Association Measures and their Combination for Arabic MWT Extraction. arXiv preprint arXiv.(2014). PP.1409-3005.

[16] Jadalla, A., &Elnagar, A. A fingerprinting-based plagiarism detection system for Arabic text-based documents. In Computing Technology and Information Management (ICCM), 2012 8th International Conference on(2012, April). (Vol. 1, pp. 477-482). IEEE.

[17] Liu, D., Li, T., & Liang, D. Incorporating logistic regression to decision-theoretic rough sets for classifications. International Journal of Approximate Reasoning, 55(1),(2014).PP. 197-210.

[18] Musa, A. B. Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics, 4(1),(2013). PP.13-24.