Journal of Information Systems and Telecommunication
Vol.10, No.3, July-September 2022, 169-179
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Corpus for Evaluation of Cross Language Text Re-use Detection Systems

Salar Mohtaj[1], Habibollah Asghari[2*]

[1]. Faculty IV, Technische University Berlin, Germany
[2]. ICT Research Institute, ACECR, Tehran, Iran

Received: 02 Feb 2022/ Revised: 06 July 2022/ Accepted: 11 July 2022

Abstract
In recent years, the availability of documents through the Internet along with automatic translation systems have increased plagiarism, especially across languages. Cross-lingual plagiarism occurs when the source or original text is in one language and the plagiarized or re-used text is in another language. Various methods for automatic text re-use detection across languages have been developed whose objective is to assist human experts in analyzing documents for plagiarism cases. For evaluating the performance of these systems and algorithms, standard evaluation resources are needed. To construct cross lingual plagiarism detection corpora, the majority of earlier studies have paid attention to English and other European language pairs, and have less focused on low resource languages. In this paper, we investigate a method for constructing an English-Persian cross-language plagiarism detection corpus based on parallel bilingual sentences that artificially generate passages with various degrees of paraphrasing. The plagiarized passages are inserted into topically related English and Persian Wikipedia articles in order to have more realistic text documents. The proposed approach can be applied to other less-resourced languages. In order to evaluate the compiled corpus, both intrinsic and extrinsic evaluation methods were employed. So, the compiled corpus can be suitably included into an evaluation framework for assessing cross-language plagiarism detection systems. Our proposed corpus is free and publicly available for research purposes.

Keywords: Cross Language Plagiarism Detection; Corpus; Text Re-Use Detection; Obfuscation.

1- Introduction

Plagiarism is the unacknowledged reuse of others' ideas or text without giving a proper credit [1]. Nowadays, due to the high availability of digital content on the web, the malpractice use of others text has been widely spread. Plagiarism detection (PD) is the act of finding patterns of text re-use between a suspicious document and source documents. With the rapid growth of documents in different languages, the increased accessibility of electronic documents, and the availability of translation tools, cross-language plagiarism has become a serious problem, and its detection requires more attention [2]. Nowadays, a vast amount of knowledge is created in rich resource languages like English, and students in low resource languages have a motivation to bring the knowledge to their language through translation. Moreover, detection of plagiarism between two pairs of languages is a more complicated task with respect to monolingual plagiarism detection (MLPD). Cross-language Plagiarism Detection (CLPD) systems try to find plagiarism cases across language pairs.

Paraphrasing and translation can be considered as connected natural language tasks. "Translation represents the preservation of meaning when an idea is rendered in words in a different language, while paraphrasing represents the preservation of meaning when an idea is expressed using different words in the same language" [3]. There are different typologies of transformation from source to the target language through translation. In other words, plagiarism between languages can occur in different types: a simple translation, translation plus paraphrasing, merging of sentences, splitting a sentence into two or more sentences in the target language, and summarization after translation.

In order to investigate various PD algorithms and evaluate their accuracy, the algorithms should be run on a plagiarism detection corpus. A PD corpus is comprised of two sets of text files named a source and suspicious documents. In order to construct a PD corpus, we should select some passages from source documents. In the next step, in order to simulate the action of plagiarism, some

✉ **Habibollah Asghari**
habib.asghari@ictrc.ac.ir

modifications should be done on the selected passages. In the final step, the paraphrased passages (henceforth called obfuscated passages) are inserted into suspicious documents. The offset and length of each passage in source and suspicious document are written into XML meta-data files.

There are some reasons that the construction of a plagiarism corpus that contains real cases of plagiarism is not a point of interest. First, because of concealed behavior of plagiarism, collecting real plagiarism cases is an expensive and time consuming task. The second reason is that using real plagiarism cases in a public domain needs consent from the original author. Third, a corpus with real plagiarism cases cannot be published due to ethical and legal issues; because of the high availability of search engines, it is difficult to anonymize the real plagiarism cases [4]. Because of the above-mentioned reasons, the researchers usually pay attention to create simulated and artificial plagiarism cases. The synthetically made plagiarized passages should be inserted into a vast amount of text data to build suspicious documents. On the other hand, the plagiarism detection algorithms should then correctly find these passages among suspicious documents and also identify their pairs in source documents. Moreover, the current researches mostly focus on creating cases of cross-language plagiarism based on sentences of parallel corpora and don't pay attention to different typologies of transformation between languages as mentioned above.

In this study, we have investigated a cross-language plagiarism detection corpus with a new approach to obfuscate the plagiarized passages. The plagiarized passages are inserted into topically related English and Persian Wikipedia articles in order to have a more realistic situation. Although we have focused our experiments on English and Persian as the source and suspicious languages, the proposed approach is not dependent on the mentioned languages.

There are some studies to construct bilingual plagiarism detection corpora from English to Hindi, Basque, Portuguese, Spanish, Hungarian, and Italian [5], [6], [7]. But they have simply used the sentences of a parallel corpus to create plagiarism fragments, and so they did not incorporate levels of obfuscation into their corpus. In other research, Asghari et al, in [15] have incorporated types of obfuscation into their corpus to build an evaluation framework. Instead, in this study, we have used a technique in selecting the plagiarism fragments from a bilingual corpus in such a way that different levels of obfuscation can be created, and so we can measure to what extent a translated plagiarized passage is hard to find. For this purpose, various factors (e.g., features based on sentence length, dictionary-based features, alignment-based features, and miscellaneous features) were used to measure the similarity of plagiarized passages.

The paper is organized as follows: In the next section, an overview of previous works in cross-language corpus construction will be presented. Our approach is described in Section 3, in which we will explain the proposed model and also the features that are used to be incorporated into the obfuscation stage. Section 4 comes with experiments and results for evaluation of the constructed corpus, including experimental setup and applying two cross-language plagiarism detection algorithms to evaluate the constructed corpus. Conclusion and recommendations for future works will be discussed in the final section.

2- Related Work

In this section, a survey on the previous research concerning the creation of cross-language plagiarism detection corpora is presented.

Using human and machine translation to translate documents from source languages to target ones can be considered as initial efforts to compile cross-language plagiarism corpora. Barrón-Cedeño et al. [1] and Pinto, Civera, Barrón-Cedeño, Juan, & Rosso [8] have used this approach to create English-Spanish and English-Italian corpora of plagiarism, respectively. In [1] five English original text fragments have been translated to Spanish by nine humans and also five automatic machine translation systems. Moreover, to evaluate plagiarism detection systems in the case of false positive detections, 46 cases of un-plagiarized fragments have been added into the corpus. The proposed corpus by Pinto et al. has been compiled by translating source English documents to 14 different plagiarized fragments using both human and machine translations [8]. Like the proposed corpus by Barrón-Cedeño et al. [1], 46 un-plagiarized fragments have been added into the corpus to simulate more realistic situations of plagiarism.

The PAN plagiarism detection corpus PAN-PC-09 is the first large-scale plagiarism detection corpus which includes a set of cross-language plagiarism cases across different language pairs [5]. The cross-language section of PAN corpus covers 10% of the corpus and includes automatically translated plagiarized fragments from German and Spanish to English. Although the monolingual part of PAN-PC-09 contains some automatic obfuscation methods to paraphrase source fragments (i.e. random text operations and semantic word variation), no obfuscation method has been used to create cross-language cases of plagiarism. Subsequent PAN-PC-10 [4] and PAN-PC-11 [9] corpora contains 14% and 11% cross-language cases of plagiarism, respectively. Moreover, to improve the quality of the cross-language corpus, 1% of automatically translated fragments of PAN-PC-11 had been corrected manually.

To create cases of plagiarism across languages, in recent research, parallel corpora have been widely used instead of incorporating human and machine translation. Parallel corpora contain several sentence pairs in two (source and target) languages, which are translations of each other [10]. ECLaPA cross-language plagiarism detection corpus has been compiled by Pereira, Moreira, and Galante [11] using the proposed methods of PAN-PC-09. This corpus is based on the Europarl Parallel Corpus and contains 300 English documents as the source and 174 Portuguese and French documents as the suspicious ones.

Potthast, Barrón-Cedeño, et al. compiled a cross-language PD corpus using JRC-Acquis parallel corpus and Wikipedia to compare the performance of different CLPD approaches across languages [12]. A total number of 23,000 parallel sentences of JRC-Acquis and 45,000 Wikipedia documents have been used to create the corpus, in which 10,000 aligned documents have been used to test the algorithms and the remaining documents have been used to train the methods.

Ceska, Toman, and Jezek created a multi-lingual plagiarism detection corpus for evaluating their proposed method for plagiarism detection based on Euro Wordnet [13]. The JRC-EU and Fairy-tale multi-lingual corpora are used for this purpose. The proposed corpus consists of 200 English reports from JRC-EU and 27 English documents of Fairy-tale as source documents and a same number of documents in Czech as the suspicious ones. Arefin et al. proposed a new approach for creating a multi-lingual plagiarism detection corpus to evaluate PD systems between Bangla and English documents [7]. They used 110 collected documents from a public university, where two groups of students have been asked to submit their reports in two different languages, namely English and Bangla. In another research, Barrón-Cedeño, Rosso, Devi, Clough, and Stevenson proposed a CL!TR task on cross-language text re-use detection across two languages: Hindi and English [6]. The participants in the competition should find potential English source documents for a Hindi suspicious one. The corpus consists of 5032 English Wikipedia documents as the source documents and 388 Hindi documents as the suspicious ones. To generate cases of plagiarism, the participants are asked to write short answers to a set of questions either by re-using the source documents or by using provided learning materials. To simulate real cases of plagiarism, they asked participants to answer questions with 4 different levels of obfuscation, including: near copy, light revision, heavy revision and no plagiarism. The last method is designed to generate answers which are not plagiarized to be used for comparison.

Although the above mentioned corpora can evaluate cross-language PD systems, they suffer from two main drawbacks:

- Lack of obfuscation degree in plagiarized fragments.
- Lack of topic similarity between plagiarized fragments and documents.

In the case of the first challenge, the reviewed corpora cannot measure the performance of plagiarism detection systems according to different levels of paraphrasing. In spite of cross-language PD corpora, the definition of obfuscation degrees has been widely used in monolingual PD corpora. In the case of second drawback, the topic similarity between plagiarized fragments and source and suspicious documents play an important role for reaching more realistic PD corpora. Ignoring topic similarity between fragments and documents can lead to detection of cases of plagiarism simply by analyzing topic drift in documents [14].

Asghari et.al., presented an approach to cross-language plagiarism detection using word embedding methods [15]. For investigating the performance of the algorithm, a corpus comprised of seven different types of obfuscation was constructed. The simulated cases of plagiarism were compiled by expert crowd workers, and the artificial ones were compiled automatically. For validation of the corpus, it was automatically checked considering the ratio of the length of plagiarized passages to the length of the documents and the distribution of plagiarized passages across the documents as well. Moreover, for evaluation of the corpus, a manual checking was done for investigating the quality of plagiarized fragments [15]. In another research from the same group, Asghari et.al., proposed a bilingual PD corpus from a sentence aligned parallel corpus. To cover different ranges of plagiarism, the degree of obfuscation has been simulated by generating plagiarized fragments by a combination of sentences with different similarity score. The corpus contains 19973 English and 7142 Persian documents [32].

In addition to the highlighted cross-lingual PD corpora, a number of monolingual corpora have been introduced in recent year, too. Al-Thwaib et.al., generated JUPlag, an Arabic PD reference corpus that is dedicated to academic language [16]. They mentioned that the corpus could be for corpus-based linguistic analyses, and also for language learning and teaching. In another research, Khoshnavataher et.al., compiled a monolingual Persian corpus from Wikipedia articles [17]. The articles have been obfuscated automatically to simulate real cases of plagiarism. We followed a similar approach to generate cross-lingual plagiarism cases in this paper. However, unlike the mentioned works, we considered the degree of obfuscation into account in a cross-lingual setting to cover a wider range of plagiarism. Briefly, our contributions in this paper are as follow:

- Construction of a large English-Persian bilingual plagiarism detection corpus, so the results of running PD algorithms on this corpus are considerably reliable.

- Incorporating paraphrasing degree into plagiarized passages. So, the similarity score of paired sentences in the corpus can be used for establishing the degree of obfuscation for plagiarism cases.
- The use of topic similarity to match between plagiarized fragments and related texts to construct suspicious documents of similar topics based on a graph clustering approach.

3- Our Approach

Our proposed approach differs from the widely used framework of previous researches for creating monolingual PD corpora. In order to construct a monolingual plagiarism detection corpus, there are three methods for creating plagiarism fragment cases, namely artificial, simulated and real approaches. As mentioned before, real cases of plagiarism are not used in PD corpora. So, the proposed methodology by Potthast et.al, [18] is a popular approach. They have used simulated and artificial methods for creating their plagiarism detection corpus. In the case of artificial plagiarism fragments, different degrees of obfuscation can be obtained by adjusting the number of operations like addition, deletion, and semantic word variation on fragments of text from a source document to be inserted as plagiarized fragments into suspicious documents. Given that the artificial method of obfuscation would generate fragments that are not understandable for humans, in this study, we proposed a new method to generate fragments with different levels of paraphrasing that are human understandable. To this end, we have used a similarity score between sentence pairs of a parallel corpus for obtaining degrees of obfuscation.

In constructing a monolingual plagiarism detection corpus, the following main steps should be accomplished:

1. Dividing documents into two distinct categories, namely source and suspicious
2. Extracting plagiarism candidate fragments from source documents
3. Applying paraphrasing methods on these fragments (contains exact copy without obfuscation, paraphrasing fragments, random shuffling, and so on)
4. Inserting obfuscated fragments into suspicious documents

Our approach to create a bilingual corpus follows the mentioned steps, except that the plagiarized fragments are extracted from a sentence aligned parallel corpus. Moreover, unlike existing bilingual PD corpora, we have investigated a degree-based paraphrasing method to better simulate real cases of plagiarism and to determine the capability of PD algorithms encountering various types of obfuscations.

For automatically constructing a cross-language corpus, in the first step, we need a parallel bilingual sentence-aligned corpus. In the next step, we should find a way to put together the paired sentences from the parallel corpus in such a way that they are topically related with each other and moreover, topically related to the surrounding text in suspicious document they are inserted. Moreover, in order to obtain some levels of obfuscation, we should apply a method of measuring the similarity between plagiarized passages.

In our bilingual PD corpus, the English and Persian Wikipedia articles have been used for the source and suspicious documents, respectively. Wikipedia is one of the largest multi-lingual corpora, which is highly popular and contains documents in different languages [19], [20]. Wikipedia is a rich vocabulary corpus that contains documents in different domains and contain a wide range of topics [21]. As a pre-processing step, the small size articles were removed from the corpus. Moreover, all the selected documents were normalized. In order to avoid instances of pseudo plagiarism passages, the near duplicate documents were removed from the data. The statistics of the documents in the corpus are presented in Table 1. In the next sub-sections we will deal with constructing the parallel corpus and also compiling the cross-language PD corpus.

Table 1: Corpus Statistics

Document Purpose	Number of Documents	27115
	% of Source Documents (English)	73%
	% of Suspicious Documents (Persian)	27%
Document Length	Short (1-500 words)	16%
	Medium (500-2500 words)	53%
	Long (2500-33000 words)	31%
	Average number of words per document	2353
	Average number of sentences per document	115
	Smallest document (by words)	300
	Largest document (by words)	32620

3-1- Construction of Plagiarism Cases

The main steps being used to construct the English-Persian cross-language plagiarism detection corpus are depicted in Figure 1. The mentioned stages in the figure to construct the plagiarism cases will be described in detail in the following sub-sections.

3-1-1- Extracting Parallel Sentences

In this step we need a parallel bilingual corpus with similarity scores for each sentence pair. Some efforts have been made in other researches that could be useful for developing a bilingual corpus. For example a method has been developed for automatic acquisition of translated web pages based on searching the hyper-links containing strings of the kind “Persian version” in order to download the versions of a given page in other languages [22].

The parallel corpus which we have presented in this paper is created by an approach that automatically extracts

parallel sentences from the web applied to English and Persian Wikipedia articles. To produce the parallel corpus, a Maximum Entropy binary classifier is trained to compute local similarities between sentence pairs of two aligned documents [19].

For building aligned paired sentences, first of all, the aligned English-Persian documents were extracted from Wikipedia. In the second step, in order to extract aligned sentences, a Maximum Entropy binary classifier has been trained in order to evaluate the local similarity between sentence pairs in Persian and English aligned documents [23]. MaxEnt classifiers are log-linear models that try to capture contextual information. They use a conditional probability of a model y given the history x and a parameter vector as follows:

$$P(y|x,\lambda) = \frac{\exp \sum_i \lambda_i f_i(x,y)}{\sum_{y' \epsilon Y} exp \sum_i \lambda_i f_i(x,y')} \quad (1)$$

Where x is the input domain which represents the history, y is a finite label set that represents the model, and $f(x,y)$ is the feature vector representation. In this equation, each feature comes with a corresponding parameter vector weight λ_i. It should be noted that MaxEnt classifiers do not depend on the correlation between features. Barrón-Cedeño, Paramita, Clough, and Rosso have investigated cross-language similarities by incorporating various features such as character n-grams, cognateness, word count ratio, and an approach based on out-links in Wikipedia pages [24]. In this research, a collection of 12 features in four categories were exploited to train the maximum entropy classifier [23]. The four categories are as follows:

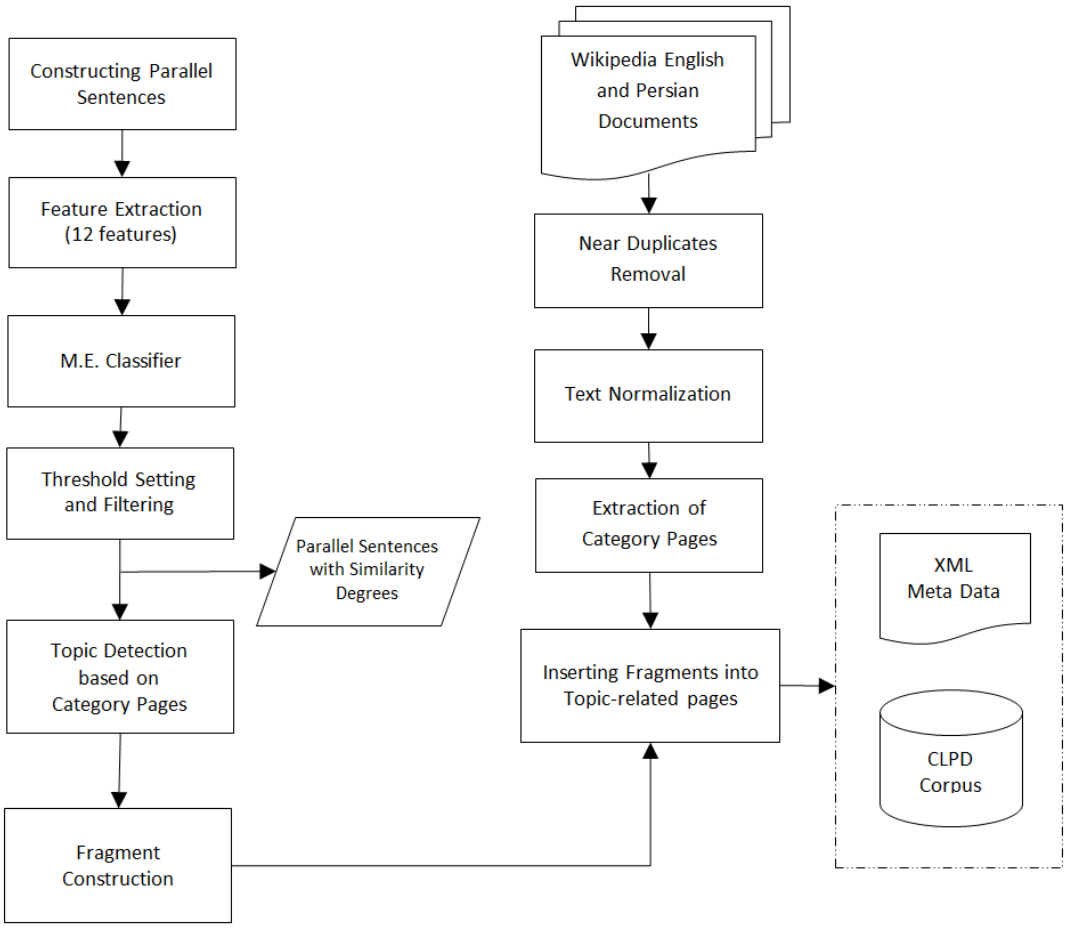

Fig. 1 Flow Diagram of Bilingual PD corpus construction

1. Features based on sentence length
2. Dictionary based features
3. Alignment-based features
4. Miscellaneous features

The features for the construction of a parallel corpus have been thoroughly described in [23]. In order to train the MaxEnt classifier, a non-parallel corpus was also constructed to compensate the bias toward the parallel sentence. As a result, by incorporating these features into the log-linear model, an alignment score is derived from the MaxEnt classifier. This score is used as a measure of similarity between paired sentences and is incorporated in the obfuscation stage of the corpus compiler.

3-1-2- Topic Extraction

In this step, the proposed approach for clustering the sentences of the parallel corpus is presented. The sentences of the parallel corpus are put together in order to construct plagiarism fragments based on their topic similarity. In addition, the plagiarism cases should be inserted into documents with similar topics. This could result in more realistic cases of plagiarism and could make the plagiarism detection process more difficult. Since the parallel aligned sentences are extracted from Wikipedia, so we used the Wikipedia rich structural properties of relevant pages to properly extract the topics and to cluster them for finding similar sentences. Among different structural features, we used the ”Categories pages” bipartite graph extracted from Wikipedia pages to separate distinct pages and extract similar ones.

For searching sentences in the Wikipedia repository, at the first stage, we extracted the relevant Wikipedia page for each sentence in the parallel corpus. Using Apache Lucene, we indexed the Wikipedia pages for finding relevant pages. In the next stage, the extracted pages are clustered based on their connectivity. The pages' connectivity was obtained from the categories- pages bipartite graph. The Infomap community detection algorithm was used to extract the clusters of pages from pages' graph [25]. Fortunato and Lancichinetti tested several community detection methods against a recently introduced class of benchmark graphs, with heterogeneous distributions of degree and community size [26]. They concluded that the Infomap method by Rosvall and Bergstrom has the best performance on the set of benchmarks they examined [26]. So, due to good performance of the Infomap community detection method, and its reliability in applications to real graphs, we used this method to extract clusters from pages' graph. In the last stage, the sentences were clustered based on the clusters of their related pages.

3-1-3- Building Plagiarism Passages

Plagiarism cases in the bilingual corpus are constructed from parallel sentences. Plagiarized fragments have been constructed from Persian sentences, and corresponding source fragments have been constructed from English sentences aligned with the source sentences.

As a result, by defining a similarity score between sentence pairs in English-Persian parallel corpus, we can make various patterns of obfuscation in plagiarized passages by putting together sentences with the same similarity scores into one plagiarized passage. Sentences with low similarity scores create high obfuscated passages, while highly similar sentences will result in low obfuscated passages. Based on this method, a combination of different plagiarism cases was built and added into the source and suspicious documents. Some examples of created fragments by sentences with different similarity scores are shown in Tables 2 and 3.

Table. 2 Example of paired sentences with similarity scores

English Sentence	Persian Sentence	Similarity Score
Willie Nelson at the Teatro	ویلی نلسون در تیترو	0.44
I hate heaven – 1998	من از آن متنفرم بهشت - ۱۹۹۸	0.51
A concise course in physics,	دوره مختصر از علم فیزیک	0.57
Economic history of ancient Greece	کتاب تاریخ کهن یونان	0.41
Sarah Bernhards – French stage and film actress	سارا برنار، هنرپیشه فرانسوی	0.53
Allen Garfield – Leo Kubelsky	الن گارفیلد – لیو کبلاسکی	0.48
President of the American oriental society	رییس جامعه شرق شناسان امریکا	0.62

Table. 3 Example of plagiarized fragments

Passage	Persian Passage	Degree of Obfuscation
President of America oriental society Sarah Bernhards – French stage and film actress	رییس جامعه شرق شناسان آمریکا، سارا برنارد، هنرپیشه فرانسوی	Low
Allen Garfield – Leo Kubelsky, I hate heaven – 1998	آلن گارفیلد – لیو کبلاسکی، من از آن متنفرم بهشت - ۱۹۹۸	Medium
Economic history of ancient Greece, A concise course in physics, Willie Nelson at the Teatro	کتاب تاریخ کهن یونان، دوره مختصر از علم فیزیک، ویلی نلسون در تیترو	High

To consider the degree of obfuscation in plagiarized fragments, we exploited a combination of sentences with different similarity scores. The similarity score of sentences in a fragment specifies the degree of

obfuscation. As shown in Table 2, for constructing “Low” obfuscated passages, we used sentences with scores between 0.5-1.0, for constructing ”Medium” obfuscated passages, we used a combination of sentences with scores between 0.5-1.0 and sentences having the scores in the range of 0.45-0.5, and for constructing ”High” obfuscated passages, a combination of sentences with scores between 0.5-1.0 and sentences having the scores of 0.4-0.45 have been used. As a result, three different degrees of obfuscation have been inserted into the bilingual corpus named as ”Low”, ”Medium”, and ”High” obfuscation which is shown in Table 4. Also, the ratio of different degrees of paraphrasing in the proposed corpus is represented in Table 5.

In this paper, the scores derived from features of the parallel sentences in parallel corpus were considered as the bases criterion for producing plagiarism cases with different level of obfuscation.

Table 4 Degree of obfuscation in plagiarism cases

Degree	Similarity scores of sentences in fragment		
	0.40 - 0.45	0.45 - 0.50	0.50 - 1
Low	-	-	100%
Medium	-	25% - 45%	55% - 75%
High	45% - 65%	-	35% - 55%

Table 5 Statistics of different degrees of paraphrasing

Plagiarism Case Statistics		
	Number of Fragments	11210
Obfuscation	Low Obfuscation	49%
	Medium Obfuscation	50%
	High Obfuscation	1%

3-2- Construction of Source and Suspicious Documents

The Persian and English articles from Wikipedia repository have been used for creating suspicious and corresponding source documents. We considered two restrictions for choosing documents. First, a length restriction was applied for choosing the documents, so that the pages with less than 300 words were not considered. Second, the chosen documents should have similar topics with parallel corpus sentences. Therefore at this point, we used the extracted clusters mentioned in the previous section for choosing proportionate pages. For this purpose, we considered the Wikipedia pages categories as the cluster label. For each cluster, the Wikipedia pages with similar categories are considered as the cluster’s pages.
After clustering pages based on their categories, they are divided into two distinct sets, namely suspicious and source ones. We considered English pages as the source and Persian pages as the suspicious ones. This is because most plagiarism cases occur from English resources translated into Persian text.

Persian belongs to Arabic script-based languages which cover Kurdish, Urdu, Arabic, Pashtu, and Persian [27]. These languages have common features such as common scripting, absence of capitalization, right to left direction, lack of clear word boundaries, and complex word structure. So, there are also some challenging issues dealing with basic NLP operations on Persian text processing, such as tokenization and stemming. We used Parsivar pre-processing toolkit for these operations [28].

3-3- Compiling the PD Corpus

In order to insert the plagiarism fragments into source and suspicious documents, the same parameters as PAN-PC-12 corpus [29] have been considered in this paper. In other words, we used the same distributions for corpus main parameters such as fragment lengths, and the number of plagiarism fragments per documents. In addition, the similarity scores derived from parallel sentences in parallel corpus were considered as the bases criterion for producing plagiarism cases with different level of obfuscation. Table 6 shows the length of plagiarism fragments in terms of the number of sentences.

The percentage of plagiarism in each suspicious document is distributed between 5% and 60% of its length. The ratio of plagiarism per suspicious documents is shown in Table 7. We have used XML tags to specify the meta-data characteristics of the plagiarized segments in suspicious documents and corresponding source documents. The corpus is tagged by specifying the offset of plagiarism cases and its equivalent fragment in both source and suspicious documents. Moreover, the length of plagiarism cases and also the degree of obfuscation (low, medium and high) have also been inserted into XML meta-data files.

Table 6 Plagiarism case statistics

Case Length	Short (20-50 words)	35%
	Medium (50-100 words)	50%
	Long (100-300 words)	15%

Table 7. Ratio of plagiarism fragments in documents

Plagiarism per Document	**Ratio**
Hardly (5% - 10%)	78%
Medium (11% - 25%)	19%
Much (26% - 60%)	3%

4- Corpus Evaluation

There are some methods for evaluation of text reuse detection corpora as follows:

- Evaluation with one or more downstream tasks to inspect the behavior of the corpus with different degrees of paraphrasing (Extrinsic evaluation).
- Evaluation with Pearson correlation coefficient to investigate how the human judgment complies with different degrees of paraphrasing in the corpus (Intrinsic evaluation)
- Evaluation of the size of the corpus; the size of the corpus is increased step by step and in each step, the corpus is evaluated with various evaluation methods. The process is stopped when the evaluation criteria doesn't change and remains fixed.
- Comparing the corpus with a standard corpus to investigate the various parameters of the corpus w.r.t. standard one.

There are also some validation experiments to validate text re-use corpora:

- A manual checking should be done for evaluating the quality of plagiarized fragments.
- The corpus can be automatically validated considering the ratio of the length of plagiarized passages to the length of the documents.
- The corpus can also be validated by inspecting the distribution of plagiarized passages across the documents.

In this section, two approaches for the evaluation of the constructed corpus have been exploited. For the first evaluation approach, the extrinsic method proposed by Clough & Stevenson has been used [30]. In this approach, we use the constructed corpus as input to a downstream task (plagiarism detection) and measure the impact of degrees of paraphrasing on the algorithm's performance. On the other hand, intrinsic evaluations directly evaluate a corpus from different point of views. In the intrinsic approach of evaluation, we have calculated Pearson correlation coefficient to investigate how the human judgment complies with different degrees of paraphrasing in the corpus.

4-1- Extrinsic Evaluation

In the case of the first approach, the main idea is that the degree of paraphrasing should affect the performance of plagiarism detection algorithms. More precisely, the algorithms' performance on finding plagiarized fragments would be decreased by increasing the degree of paraphrasing. Clough and Stevenson have used a simple n-gram based method as a baseline for measuring the influence of the obfuscation degree in the performance of a plagiarism detection method [30]. In this research, we have applied machine translation plus monolingual analysis (T+MA) and Latent Semantic Analysis (LSA) methods along with the Vector Space Model (VSM) to determine whether they can distinguish between the various levels of paraphrasing. In order to test the validity of our approach and compare the results, we evaluated the performance of the methods on the total corpus and also various parts of the corpus with different degrees of paraphrasing as follow:

- Total corpus
- Low degree of paraphrasing part of the corpus
- Medium degree of paraphrasing part of the corpus
- High degree of paraphrasing part of the corpus

For T+MA method, the Targoman [1] machine translation API was used for translating documents in Persian into English. Targoman is the state-of-the-art English-Persian machine translation system that is freely available to be used. After translating Persian documents to English, the source and suspicious documents have been compared to detect cases of plagiarism. To this end, the cosine similarity between vectors (VSM model based on tf-idf weighting) of both source and suspicious sentences have been computed using the following equation:

$$Cos\theta = \frac{S_{src}.S_{susp}}{\|S_{src}\|.\|S_{susp}\|} \quad (2)$$

Where $S_{src}.S_{susp}$ is the dot product of the source and suspicious sentences' vectors, and $\|S_{src}\|.\|S_{susp}\|$ are norms of the source and suspicious sentences, respectively. The Plagdet measure was introduced by Potthast et.al, for evaluation of the algorithm's performance against different levels of paraphrasing in the corpus [18]. Plagdet is a common metric for evaluating PD systems which is a weighted F-measure as depicted in the following Equations:

$$Plagdet\ (S,R\ ) = \frac{F_1}{1 + gran(S,R)} \quad (3)$$

$$Prec\ (S,R\ ) = \frac{1}{|R|}.\sum_{r\in R}\frac{U_{s\in S}\ (s \cap r)}{|r|} \quad (4)$$

$$Recall\ (S,R\ ) = \frac{1}{|S|}.\sum_{s\in S}\frac{U_{r\in R}(s \cap r)}{|s|} \quad (5)$$

$$F_1 = 2.\frac{Precision.Recall}{Precision + Recall} \quad (6)$$

Where S denote the set of plagiarism cases in the suspicious documents of the corpus, and R denote the set of plagiarism that are detected by the detector for these documents, and F_1 denotes the $F-Measure$. Moreover, the parameter $gran$(S,R) indicates the one-to-one correspondence between detected and desired source fragments.

[1] www.targoman.com

The performance of detection in diff erent parts of the corpus is depicted in Figures 4 and 5. The obtained results of applying T+MA and LSA methods on the proposed corpus show the influence of levels of obfuscation on algorithms' performance in detecting cases of plagiarism in the corpus regardless of chosen parameters.

4-2- Intrinsic Evaluation

In addition to the extrinsic evaluation of the proposed corpus, we have used an intrinsic approach in which the human judgment has been used as an assessment criterion to investigate how human degrees comply with fragments' degrees of paraphrasing, using the *Pearson correlation coefficient*. For this purpose, the approach by Paramita, Clough, Aker, & Gaizauskas has been used [31]. A collection of eight Persian speaking persons fluent in reading English texts were asked to assess the similarity of English and Persian fragments based on a 9-point Likert scale in the range of 1 to 9 (very low similarity to very high similarity). A total number of 150 fragments have been annotated by at least 3 persons in the mentioned range of similarity. The Pearson correlation coefficient has been used to compute the degree of correlation between automatically computed degrees of paraphrasing and human judgments based on the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

Where x_i and y_i are "automatically generated degrees" and "degrees assigned by human judgments", respectively. Moreover, $\bar{x}$ and $\bar{y}$ are the mean of degrees.

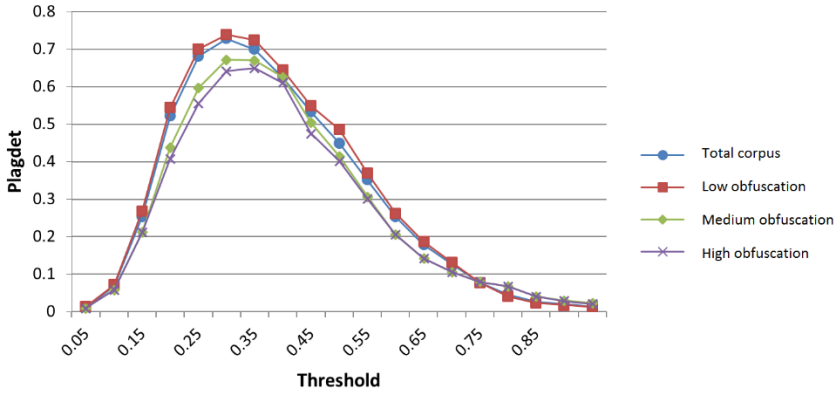

Fig. 4 Changes of Plagdet vs. cosine similarity by applying T+MA on various parts of the corpus

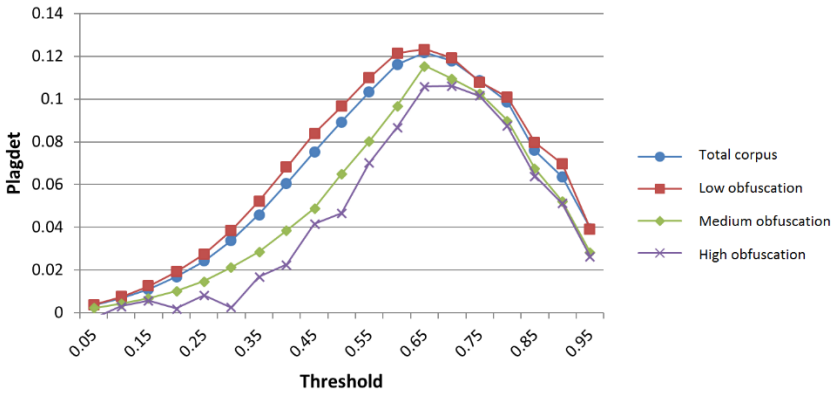

Fig. 5 Variation on Plagdet vs. cosine similarity threshold, by applying LSA on various parts of the corpus

The results show a *Pearson correlation coefficient* of r = 0.665 which shows that; there is a significant correlation between the paraphrasing degrees assigned by the assessors and automatically assigned degrees, that is higher than related studies like what has been done in [31].

To conclude, we applied two different approaches to evaluate the proposed corpus based on level of paraphrasing. Our results show that the proposed corpus and the applied method for ranking the level of

paraphrasing is accurate in both extrinsic and intrinsic evaluations.

5- Conclusions and Future Works

In this study, we have constructed an English-Persian bilingual plagiarism detection corpus by exploiting Wikipedia English and Persian articles as main resource data for source and suspicious text, respectively. We also used a parallel bilingual corpus to construct plagiarized passages. In order to create passages with various degrees of obfuscation, we incorporated some global and local features in order to measure the similarity between plagiarized sentences in the source and target languages. As a result, we can measure the degree of obfuscation in the aforementioned plagiarism detection corpus. So we can adjust the complexity of obfuscation in plagiarized passages in the dataset.

In order to build a more realistic corpus, the plagiarized passages were inserted into the topically related source and suspicious text. We applied two different PD algorithms on the bilingual corpus as extrinsic evaluation and also human judgment evaluation as intrinsic evaluation approach. The results prove the validation of the proposed obfuscation method in our bilingual corpus. The constructed CLPD corpus is freely available on the web for research purposes1.

Further improvements can be conducted by adding compositional text where one sentence in the source document is translated and converted into two or more sentences in the suspicious document or vice versa. Moreover, different types of obfuscation can be applied to the corpus as well. So we can reach a multi-type multi-degree obfuscation CLPD corpus.

Acknowledgments

This research was supported by the ICT Research Institute affiliated to Academic Center for Education Culture and Research (ACECR). We thank our colleagues from this institution who provided the expertise that greatly assisted us in this research. Our special thanks go to Dr Heshaam Faili for his valuable help along the way which greatly assisted this research.

References

[1] A. Barrón-Cedeño, P. Rosso, D. Pinto, and A. Juan, "On Cross-lingual Plagiarism Analysis using a Statistical Model", Proceedings of the ECAI'08 workshop on uncovering plagiarism, authorship and social software misuse, Patras, Greece, 22 July 2008 (Vol. 377). CEUR-WS.org.

[2] N. Ehsan, and A. Shakery, "Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information", *Information Processing and Management, vol.* 52, no. 6, pp. 1004-1017, 2016.

[3] C. Callison-Burch, "Paraphrasing and translation", Doctoral Dissertation, School of Informatics, University of Edinburgh, 2007.

[4] M. Potthast, A. Barrón-Cedeño, A. Eiselt, B. Stein, and P. Rosso, "Overview of the 2nd international competition on plagiarism detection". In CLEF 2010 labs and workshops, notebook papers, 22-23 September 2010, Padua, Italy (Vol. 1176). CEUR-WS.org.

[5] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso, "Overview of the 1st international competition on plagiarism detection". In 3rd PAN workshop; Uncovering plagiarism, authorship and social software misuse (PAN 09), San Sebastian, Spain, 10 September 2009, pp. 1–9.

[6] A. Barrón-Cedeño, P. Rosso, S. L. Devi, P. D. Clough, and M. Stevenson, "PAND@FIRE: Overview of the cross-language !ndian text re-use detection competition." Multi-lingual information access in south asian languages - second international workshop, FIRE 2010, gandhinagar, india, february 19-21, 2010 and third international workshop, FIRE 2011, Bombay, India, 2-4 December 2011, revised selected papers (Vol. 7536, pp. 59–70). Springer.

[7] M. S., Arefin, Y. Morimoto, and M. A. Sharif. "BAENPD: A Bilingual Plagiarism Detector", *Journal of Computers.* vol. 8, no. 5, pp. 1145-1156, 2013.

[8] D. Pinto, J. Civera, A. Barrón-Cedeño, A. Juan, and P. Rosso. "A statistical approach to cross-lingual natural language tasks" *Journal of Algorithms*, vol 64, no. 1, pp. 51-60, 2009.

[9] M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein, B., and P. Rosso, "Overview of the 3rd international competition on plagiarism detection". In CLEF 2011 labs and workshop, notebook papers, 19-22 September 2011, Amsterdam, the Netherlands (Vol. 1177). CEUR-WS.org.

[10] W. A. Gale, and K. W. Church, "A program for aligning sentences in bilingual corpora." *Computational Linguistics*, vol. 19, no. 1 pp. 75-102. 1993.

[11] R. C. Pereira, V. P. Moreira, and R. Galante, "A new approach for cross-language plagiarism analysis". Multi-lingual and multimodal information access evaluation: International conference of the cross-language evaluation forum, CLEF 2010, Padua, Italy, 20-23 September 2010. Proceedings (Vol. 6360, pp. 15–26). Springer.

[12] M. Potthast, A. Barrón-Cedeño, B. Stein, and P. Rosso, "Cross-language plagiarism detection", *Language Resources and Evaluation*, vol. 45, no. 1, pp. 45–62, 2011.

[13] Z. Ceska, M. Toman, and K. Jezek, "Multi-lingual plagiarism detection". In 13th international conference on Artificial intelligence: Methodology, systems, and applications, (AIMSA 2008), Varna, Bulgaria, September 4-6, 2008. Proceedings (Vol. 5253, pp. 83–92). Springer.

[14] M. Potthast, M., Hagen, M., Völske, M. and B. Stein, "Crowdsourcing interaction logs to understand text reuse from the web", In Proceedings of the 51st annual meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), Sofia, Bulgaria, 4-9 August 2013, pp. 1212-1221.

[15] H. Asghari, O. Fatemi, S. Mohtaj, H. Faili, and P. Rosso. "On the use of word embedding for cross language plagiarism detection", *Intelligent Data Analysis*, vol. 23, no. 3, pp. 661-680, 2019.

[1] www.ictrc.ac.ir/corpus/bilingual_persian_english_corpus(hamta3).zip

[16] E. Al-Thwaib, B. H. Hammo, and S. Yagi, "An academic Arabic corpus for plagiarism detection: Design, construction and experimentation", *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp.1-26. 2020.

[17] K. Khoshnavataher, V. Zarrabi, S. Mohtaj, S., and H. Asghari, "Developing monolingual Persian corpus for extrinsic plagiarism detection using artificial obfuscation", Notebook for PAN at CLEF 2015.

[18] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso. "An evaluation framework for plagiarism detection", In *COLING 2010: 23rd International Conference on Computational Linguistics, 23-27 August 2010, Beijing, China*, posters volume, pp. 997-1005.

[19] S. F. Adafre, and M. De Rijke, "Finding similar sentences across multiple languages in Wikipedia". In Proceedings of the 11th conference of the European chapter of the Association for Computational Linguistics, 4 April 2006, Trento, Italy, pp. 62–69

[20] P. G. Otero, and I. G. L´opez, "Wikipedia as multi-lingual source of comparable corpora", *In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pp. 21–25, 2010,

[21] T. Wang, R. Di, and J. Song, "A Novel Online Encyclopedia-Oriented Approach for Large-Scale Knowledge Base Construction", *J. Softw.*, vol. 9, no 2, pp. 482–489, 2014.

[22] P. Resnik, "Mining the web for bilingual text", In Proceedings of the 27th annual meeting of the Association for Computational Linguistics (ACL), university of Maryland, College Park, Maryland, USA, 20-26 June 1999. pp. 527-534.

[23] H. Zamani, H. Faili, A. Shakery, "Sentence alignment using local and global information", *Computer Speech & Language*, 39, pp. 88-107, 2016. doi: 10.1016/j.csl.2016.03.002

[24] A. Barrón-Cedeño, M. L. Paramita, P. D. Clough, and P. Rosso, "A comparison of approaches for measuring cross-lingual similarity of Wikipedia articles". Advances in information retrieval - 36th European Conference on IR Research, (ECIR 2014), Amsterdam, the Netherlands, 13-16 April 2014. Proceedings (Vol. 8416), pp. 424–429, Springer.

[25] M. Rosvall, and C. T. Bergstrom, C. T. "Maps of random walks on complex networks reveal community structure". Proceedings of the National Academy of Sciences of the USA, 105(4), 2008, pp. 1118–1123.

[26] S. Fortunato, and A. Lancichinetti, "Community detection algorithms: A comparative analysis: invited presentation, extended abstract. In 4th international conference on performance evaluation methodologies and tools, VALUETOOLS'09, Pisa, Italy, 20-22 October 2009, pp. 1-2. ICST/ACM.

[27] A. Farghaly, "Computer processing of Arabic script-based languages: current state and future directions". In Proceedings of the workshop on computational approaches to Arabic script-based languages, Stroudsburg, PA, USA, 28 August 2004, pp. 1-1.

[28] S. Mohtaj, B. Roshanfekr, A. Zafarian, and H. Asghari. "Parsivar: A language processing toolkit for Persian", In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 7-12 May 2018, Miyazaki, Japan, pp. 1112-1118,

[29] M. Potthast, T. Gollub, M. Hagen, J. Kiesel, M. Michel, M., A. Oberlander, B. Stein, B. "Overview of the 4th international competition on plagiarism detection". In CLEF 2012 evaluation labs and workshop, online working notes, Rome, Italy, 17-20 September 2012 (Vol. 1178). CEUR-WS.org.

[30] P. Clough and M. Stevenson, "Developing a corpus of plagiarized short answers", Language Resources and Evaluation, vol. 45, no 1, pp. 5–24, 2011.

[31] M. L. Paramita, P. D. Clough, A. Aker, A., and R. J. Gaizauskas. "Correlation between similarity measures for inter-language linked Wikipedia articles". In Proceedings of the eighth international conference on language resources and evaluation, (LREC 2012), Istanbul, Turkey, 23-25 May 2012, pp. 790–797.

[32] H. Asghari, K. Khoshnava, O. Fatemi, and H. Faili, "Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus", Notebook for PAN at CLEF, 2015.