IJICTR
International Journal of Information & Communication Technology Research

Volume 17- Number 2 – 2025 (59-74)

Multi-type Obfuscation Corpus for Cross-Lingual Plagiarism Detection

Habibollah Asghari*
Department of Advanced Information Systems
ICT Research Institute (ACECR)
Tehran, Iran
habib.asghari@ictrc.ac.ir

Salar Mohtaj
Speech and Language Technology (SLT) Department
German Research Centre for Artificial Intelligence
(DFKI), Labor Berlin, Berlin, Germany
salar.mohtaj@dfki.de

Received: 3 April 2024 – Revised: 27 March 2024 - Accepted: 2 April 2024

Abstract—**In recent years, due to the high availability of documents through the Internet, plagiarism is becoming a serious issue in many fields of research. Moreover, the availability of machine translation systems facilitates the re-use of textual content across languages. So, the detection of plagiarism in cross-lingual cases is now of great importance especially when the source and target language are different. Various methods for automatic detection of text reuse have been developed whose objective is to help human experts investigate suspicious documents for plagiarism cases. For evaluating the performance of theses plagiarism detection systems and algorithms, we need to construct plagiarism detection corpora. In this paper, we propose an English-Persian plagiarism detection corpus comprised of different types of paraphrasing. The goal is to simulate what would be done by humans to conceal plagiarized passages after translating the text into the target language. The proposed corpus includes seven types of paraphrasing methods that cover (but not limited to) all of the obfuscation types in the previous works into one integrated CLPD corpus. To evaluate the corpus, an extrinsic evaluation approach has been applied by executing a wide variety of plagiarism detection algorithms as downstream tasks on the proposed corpus. The results show that the performance of the algorithms decreases by increasing the obfuscation complexity.**

Keywords: Cross-lingual plagiarism detection, Corpus construction, Obfuscation strategy, Translation obfuscation

Article type: Research Article

Publisher: ICT Research Institute

I. INTRODUCTION

Plagiarism is defined as appropriating others' words or intellectual property without providing proper citation to them. In other words, plagiarism is the task of unacknowledged reuse of others' ideas or texts without giving proper credit or permission. Nowadays, due to the high availability of digital content on the internet, the re-use of others' text without giving a proper credit has been widely spread. Moreover, the increased accessibility of electronic documents, the rapid growth of documents in different languages, and the availability of automatic translation tools, cross-language plagiarism has become a serious problem in the field of academic integrity and its detection requires more attention [1]. Cross-lingual plagiarism detection (CLPD) systems try to find the

* Corresponding Author

plagiarism cases across language pairs. The challenge of cross-lingual plagiarism detection is more serious when the plagiarist tries to paraphrase the text after translation.

In order to evaluate the performance of various plagiarism detection (PD) algorithms, they should be tested against a plagiarism detection corpus. When the algorithm is run on the corpus, the system must determine whether a suspicious document has passages taken from the source documents in another language. Moreover, it should accurately determine the off set and the length of the plagiarized passages inside the suspicious document.

There are three diff erent approaches to construct a plagiarism detection corpus. In the first approach, real cases of plagiarized documents can be used to build the corpus. The second approach comes with generating plagiarized passages using human crowds to simulate cases of plagiarism. The last approach comes with automatically generating the artificial plagiarized passages.

There are certain reasons that compiling a real plagiarism detection corpus is not a point of concern. First, because of concealed behavior of plagiarism, collecting real plagiarism cases is very time-consuming and costly. In addition, the use of real plagiarism cases in the public domain requires the consent of the original author [2]. Therefore, the researchers usually are often interested in creating simulated and artificial plagiarism cases. The synthetically made plagiarized passages must be inserted into a large amount of textual data to compile suspicious documents. The plagiarism detection algorithms should correctly find these passages among the suspect documents and also identify the corresponding pairs in the source documents.

In this study, we have developed an English-Persian cross-lingual corpus for detailed comparison task of plagiarism detection based on a new approach for obfuscating the plagiarized passages. Moreover, in order to bring the corpus under a more realistic situation, we have inserted the plagiarized passages into topically related text documents. We also have exploited a more sophisticated strategy using various types of paraphrasing to cover diff erent types of obfuscation. It should be also mentioned that, although we have focused our experiments on English and Persian as source and target languages, the proposed approach is not restricted to the mentioned languages and can be extended to other language pairs.

Our paper is organized as follow: In section 2, an overview of previous work on cross-lingual corpus construction will be discussed. Our approach is presented in Section 3, in which we will discuss the proposed model and also the features that have been used for incorporating into the obfuscation stage. In section 4, the steps toward the construction of the corpus will be described in detail. Section 5 deals with experiments and results for the evaluation of the constructed corpus. Finally, in the last section conclusion and recommendations for future works will be discussed.

II. Related Work

Research for construction of cross-lingual plagiarism detection corpora was started in 2008. In a Czech-English CLPD system proposed in [5], a new method called MLPlag have been investigated using EuroWordNet thesaurus. For their experiments, they constructed two distinct multilingual corpora. The first corpus (JRC-EU) is composed of 400 texts randomly selected from European Union legislative documents. It contains 200 reports written in English and the corresponding texts in Czech. The second corpus (Fairy-tale) contains a smaller set of text documents with a simplified vocabulary. The corpus is composed of 54 documents, half of them in English and the remaining in corresponding translations in Czech.

The PAN plagiarism detection corpus PAN-PC-09 that was introduced in [6], includes a set of cross-lingual plagiarism cases across different language pairs. The cross-lingual section covers about 10% of the whole corpus and includes automatically translated plagiarized fragments from German and Spanish to English. The corpus is based on public domain book-length documents from the Project Gutenberg. The monolingual part of the PAN-PC-09 exploits some methods for automatic obfuscation to paraphrase the source fragments (such as semantic word variations and random text operations). Moreover, the translation has also been used as an artificial obfuscation method to create cross-lingual fragments. The PAN-PC-10 [7] with 27073 documents and 68558 plagiarism cases is a subsequent of PAN-PC-09 corpus which contains about 14% of cross-lingual plagiarism cases. In the third international competition on plagiarism detection, a revised version of the previous PAN corpora has been introduced [8]. About 11% of the corpus is cross-lingual Dutch-English and Spanish-English documents. In comparison to the previous versions of the corpus, it has a significantly larger portion of plagiarism that is obfuscated by translation, translation plus paraphrasing, and the addition of manually translated plagiarism. These changes were done because in the previous version of the corpus, automatically translated cases of plagiarism could be easily detected using machine translation APIs.

Pinto et.al, [9] have proposed a corpus by translating source English documents to Italian plagiarized fragments using both human translation and machine translation tools as well. Moreover, 20 un-plagiarized fragments were added into the corpus to simulate more realistic situations of plagiarism.

Potthast et al. in [10] have compiled a cross-language PD corpus in six languages to evaluate different cross- lingual plagiarism detection algorithms. The corpus includes German, Spanish, French, German, and Polish languages, as well as English as the source language. About 120 thousand documents from JRC-

Acquis parallel cross- lingual corpus as well as Wikipedia articles were used in the construction of this corpus. These documents have been selected in such a way that for each test document, there exist documents with high similarity in all the above six languages. The JRC-Acquis parallel corpus contains legal texts of EU documents translated and aligned into 22 EU languages. Those documents which contain aligned versions of all the aforementioned languages were considered as the plagiarized part of the compiled corpus.

For evaluating the proposed method of plagiarism detection investigated in [3], a corpus has been constructed named ECLaPA, which is composed of two corpora. The first corpus contains monolingual plagiarism cases and the other contains multilingual plagiarism cases. Both corpora contain exactly the same plagiarism cases. In the multilingual corpus, the suspicious documents are written in English, whereas the source documents are written in Portuguese or French. The ECLaPA has been created based on the Europarl parallel corpus. Of the 300 suspicious documents in each corpus, 100 of them did not contain plagiarism cases. Also, of the 348 source documents in each corpus, 100 of them were not used as source of plagiarism. Each corpus has a total of 2169 plagiarism cases; about 30% are short passages (less than 1500 characters), 60% are medium passages (from 1501 to 5000), and 10% are large passages (from 5001 to 15000). The suspicious passages have been selected randomly from Portuguese or French documents, and the equivalent English passages have been inserted into an English document.

In a PAN-FIRE shared task on Indian-English plagiarism detection, a corpus for cross-lingual text re-use between English and Hindi has been manually constructed [11]. This task was document level; i.e. no specific fragments inside the documents were expected to be identified. The corpus includes a total of 5,032 English Wikipedia articles with topics in computer science and tourism and about 388 documents written in Hindi. A set of simulated plagiarized documents was created by crowd workers. Participants were provided a set of questions and they were asked to write a short answer, either by re-using text from Wikipedia or by looking at learning material from textbooks, lecture notes, and so on. To simulate different obfuscation degrees, the participants were asked to write the answer using one of the four methods of paraphrasing. In the first method (near copy), the participants were asked to answer the question by copying text from the relevant Wikipedia articles using machine translation tools. In the second method (Light revision), the participants were asked to base their answers on text that is found in Wikipedia articles with simple paraphrasing. The participants were allowed to use machine translation tools. In the third method, (Heavy revision), participants were asked to base their answer on relevant Wikipedia articles and rephrase the text with different wordings and structure to generate an answer with the same meaning as the source text,. They were not allowed to use automatic translation tools. In the fourth method (No plagiarism), participants were provided with learning materials in the form of lecture notes, textbooks, or web pages to answer the relevant question. Participants were asked to read these materials and then attempt to answer the question using their own knowledge, as well as what they learned from the provided materials.

Researchers in [12] have investigated a cross-lingual English-Indonesian plagiarism detection system to examine different pre-processing tasks on the performance of the system. The plagiarized passages have been generated by literal translation. The corpus contains English documents on some limited topics. The corpus was divided into four sections. The sections are constructed from few plagiarized sentences up to whole plagiarized documents.

A cross-lingual plagiarism detection corpus which is comprised of German-English and Hungarian-English cases of plagiarism has been proposed in [13]. The corpus contains very small 100-sentence long manually translated Hungarian passages from Wikipedia to evaluate the proposed methods for CLPD. The English Wikipedia and a parallel corpus containing 65000 parallel sentences from *Wikipedia* in the original English, translated Hungarian and translated German are used to compile the corpus. Google Translate API has been used to generate cases of cross-lingual text reuse from English sentences. Moreover, two dictionaries (English-Hungarian with 700,000 word pairs, and English-German with 150,000 word pairs) were used to create plagiarism cases.

The first Bangla-English PD corpus has been compiled in [14] with a total number of 110 documents for their experiments. Among 110 documents, 50 Bangla documents and the corresponding 50 English documents were used as training documents. The remaining 10 documents were used for test purposes. These documents were collected from a department of a public university. Students were asked to submit their reports individually from a specific domain. A total number of 110 students were divided into two groups; one group submitted their reports in Bangla and the other group submitted their reports in English.

In the PAN 2015 shared task, an English-Persian corpus for the task of "Text alignment corpus construction" was proposed by [15]. To compile cases of plagiarism across languages, the approach has exploited sentences from an English-Persian parallel corpus. The Wikipedia documents were used for constructing the main body of source and suspicious documents. Moreover, a parallel English-Persian sentence-aligned corpus was exploited to construct the plagiarized passages. The cases of plagiarism have been constructed using a parallel corpus. The plagiarized fragments of suspicious document have been constructed from Persian sentences and the corresponding source fragments were constructed from English sentences. To consider obfuscation degrees in the plagiarized fragments, a

combination of sentences with different similarity scores were selected. So, the number of sentences as well as their similarity scores in a fragment, specifies the degree of obfuscation in a fragment. Four different degrees of obfuscation were defined based on similarity scores. By inserting the constructed plagiarized passages into the documents with related topics, a cross-lingual English-Persian plagiarism detection corpus was established. Although a number of works have been accomplished in mono-lingual plagiarism detection [16], less attention has been paid to algorithms and corpora in cross-lingual domain.

In the research conducted by Hanif et al., an Urdu-English cross-language evaluation corpus called CLUE has been developed to evaluate plagiarism detection methods [17]. The collected documents have been selected from the resources available on the web. The documents were in the subject area of computer science and general articles. Fragments of source texts were collected from Wikipedia and divided into three categories; short length items (less than 50 words), medium length items (between 50 and 100 words) and long length (between 100 and 200 words). In order to create simulated text fragments, some university students were asked to rewrite Urdu text sources and create plagiarized fragments in English. Three methods have been used to create these passages: In the Near Copy method, the participants are asked to use machine translation tools to produce plagiarized passages. In the Light Revision method, the source text passages are automatically translated from Urdu to English, and then the translated text passages are referred to an automatic text paraphrasing tool, so that finally the rewritten text pieces are obtained in English. In the Heavy Revision method, the participants are asked to manually translate Urdu texts into English. Out of 500 suspicious documents, plagiarized fragments were inserted into 270 topically related documents and no changes were made in the remaining 230 documents. [17]

A multi-lingual, multi-style and multi-granularity plagiarism detection corpus is compiled in [18]. To allow a rigorous evaluation of the state-of-the-art methods, the corpus was created with multiple granularities of aligned textual units (e.g. sentence- and chunk- level). Regarding the resources used, the corpus consists of both human and machine translation fragments. Some of the previously used resources like JRC-Acquis corpus, Europarl corpus, Wikipedia collections, and Webis-CLS-10 have been reused to construct this corpus. Moreover, to enrich these corpora, a collection of documents from PAN-2011 and conference papers have been added to the resources. To evaluate the corpus, a manual check has been performed on more than 1,300 randomly chosen aligned chunks, providing an alignment confidence greater than 92%.

An English-Urdu cross-lingual corpus English-Urdu PD corpus (CLEU) has been developed by Muneer et al for the purpose of evaluating plagiarism detection systems [19]. This corpus contains 3235 pairs of English-Urdu passages. The source data is extracted from the collection of English news agencies and Urdu data extracted from Urdu newspapers and it contains real cases of plagiarism (text reuse). The cases of plagiarism of this text are classified into three categories: exact copy, paraphrased copy, and independently written (the meaning of the two texts is the same, but not necessarily copied). The texts were marked by three computer students.

In order to investigate word embedding algorithms in plagiarism detection, a corpus was compiled to measure the performance of the bilingual word embedding algorithms against previous ones [20]. This corpus has various types of paraphrasing passages.

A large Urdu-English cross-lingual plagiarism detection corpus has been developed by Haneef, et al. [21]. The corpus includes 2395 pairs of documents (540 automatic translations, 239 artificial obfuscation, 508 manual obfuscation, and 808 documents without plagiarism cases). A linguistic analysis has also been done on the plagiarized fragments.

In a research investigated for compiling an English-Persian cross-language plagiarism detection corpus, the researchers have exploited parallel bilingual sentences and artificially generate passages with various degrees of paraphrasing [22]. To achieve more realistic text documents, the plagiarized passages have been inserted into topically related English and Persian Wikipedia articles.

In an approach to English-Arabic cross lingual plagiarism detection, a novel method is used named CL-CTS-CBOW to improve the textual similarity of the approach, and moreover, a method called CL-WES for improving the syntax features was exploited. Afterward, the approach has been improved by the IDF weighting method [32]. They have used four Arabic-English corpora, comprised of books, Wikipedia, EAPCOUNT, and MultiUN, which have more than 10,017,106 sentences with supported parallel and comparable assemblages which conceals several subjects.

In another research for the task of cross lingual plagiarism detection, they have introduced a new multilingual retrieval model named as Cross-Language Ontology-Based Similarity Analysis (CL-OSA) [33]. The model represents documents as entity vectors obtained from the open knowledge graph Wikidata. For compiling the corpus, they have selected 2,000 aligned documents from each of the following five corpora, using random sampling:

- PAN-PC-11 corpus. For the candidate retrieval evaluation, the test cases were sampled from the 2,921 Spanish-English aligned document pairs in the corpus.
- ASPEC-JE: A subset of the Asian Scientific Paper Excerpt Corpus (ASPEC) which contains abstracts

of about two million research papers that were manually translated from Japanese to English.

- ASPEC-JC: A subset of the Asian Scientific Paper Excerpt Corpus (ASPEC) which contains paragraphs from the research papers that were manually translated from Japanese to Chinese.
- JRC-Acquis: A subset of legislative texts of European Union's Joint Research Centre (JRC) in 22 languages. which sampled from the 10,000 document pairs in the English-French subset of the corpus.
- Europarl: A subset of European Parliament proceedings in 21 European languages, that sampled from the 9,443 document pairs in the English-French subset of the corpus.

The proposed model for English–Hindi cross-lingual plagiarism detection, combines the convolutional neural network (CNN) and bidirectional long short-term memory (Bi-LSTM) network to learn the semantic similarity among the language pairs, in which the CNN model learns the local context of words, whereas the Bi-LSTM model learns the global context of sentences [34]. For evaluating the performances of the proposed model, Microsoft paraphrase corpus was converted into English–Hindi language pairs.

The researchers in [35] have investigated the problem of cross-lingual plagiarism in academic works of European universities. The system composes the methods of statistical machine translation and deep learning methods based on the contextualized word embeddings, such as BERT and its multilingual version, LaBSE. To analyze the efficiency of the proposed method, they have used two corpora. The first one is a synthetic dataset generated using machine translation systems for three language pairs including English-Russian, Italian-German, and Swedish-Czech. The source documents randomly sampled from Wikipedia. The second corpus is based on real theses obtained from the repository of open access theses and dissertations (https://oatd.org). A real dataset comprised of 10202 academic graduation theses in five languages (German, French, Portuguese, Spanish, and Swedish) were analyzed. As an external collection (source documents) they have used a set of 50 million scientific documents including documents from open web resources and also the papers available in the Wiley online library.

In a research investigated by Zubarev et al. for Russian-English cross-lingual plagiarism detection, for the evaluation of the task, they automatically translated Paraplag (monolingual dataset for plagiarism detection) from Russian to English [36]. Paraplag dataset contains manually written essays on the given topic. Authors of essays should have used at least five sources, which they had to search by themselves when composing essays. They have used 300 sentences from WMT-News dataset, 300 sentences from News-Commentary dataset, and 300 sentences from cross-lingual essays. Two types of data exist, the first one consists of copy-paste or moderately disguised essays, whereas the second one contains only heavily disguised essays.

The general information of all of the above-mentioned corpora is presented in Table 1. Unlike mono-lingual plagiarism detection corpora, the proposed cross-lingual corpora do not take into account the variety of obfuscation. In other words, the mentioned corpora cannot simulate real situation of plagiarism from the paraphrasing point of view or type of obfuscation.

In this paper, we have tackled with the problem of cross-lingual corpus construction by incorporating various types of obfuscation into the corpus considering topic similarity between documents. Shortly our contribution is as follows:

- The proposed corpus includes seven types of paraphrasing methods that cover (but are not limited to) all of the aforementioned obfuscation types in the previous works into one integrated CLPD corpus.
- Cross-lingual topic modeling of source and suspicious documents and plagiarized fragments. The plagiarized fragments are inserted into topically related source and suspicious documents. This results in construction of a more realistic corpus.
- Applying a wide variety of plagiarism detection algorithms on the proposed corpus as extrinsic evaluation.

The constructed cross-lingual plagiarism detection corpus is made available for the research purposes on the Web[1].

III. Our Approach

In this section, we describe the proposed approaches to obfuscate the passages and generate cases of plagiarism in details.

Potthast et al. introduced 3 levels of plagiarism authenticity; real, simulated, and artificial plagiarism [2]. Since publishing real cases of plagiarism can lead to legal problems, all of the PD corpora try to simulate what would be done in a real scientific theft instead of using real plagiarism cases. The simulated strategies to create plagiarism detection corpora try to emulate real plagiarism using crowd-workers. In this scenario, the crowds are asked to rewrite some text passages in such a way that the rewritten version has the same meaning as the original, but with a different wording. Same pattern can be used to simulate cross-lingual plagiarism by asking crowds to translate and paraphrase passages from source language to the target one. Although the simulated approach can generate cases of plagiarism very similar to real ones, but it is costly in terms of both human resources.

[1] http://www.ictrc.ac.ir/corpus/HAMTA-CL.rar

TABLE I. STATISTICS OF CLPD CORPORA

Reference	Corpus Name	Src Lang[a]	Sus Lang[b]	Obf. type[c]
Ceska et al. (2008) [5]	JRC-EU	English	Czech	No obf.
Potthast et al. (2009) [6]	PAN-PC-09	English	German Spanish	Artificial
Potthast et al. (2010) [7]	PAN-PC-10	English	German Spanish	Artificial
Potthast et al. (2011) [8]	PAN-PC-11	English	German Spanish	Artificial Simulated
Pinto et al. (2009) [9]	-	English	Italian	Artificial Simulated
Pereira et al. (2010) [3]	ECLaPA	English	Portuguese French	Artificial
Barrón-Cedeño et al. (2011) [11]	CL!TR	English	Hindi	Simulated
Potthast, et al. (2011) [10]	-	English	Spanish German French Dutch Polish	No Obfuscation
Alfikri et al. (2012) [12]	-	English	Indonesian	Literal Translation
Pataki (2012) [13]	-	English	German Hungarian	Artificial Simulated
Areffin et al. (2013) [14]	-	English	Bangla	Simulated
Hanif et al. (2015) [17]	CLUE	Urdu	English	Simulated
Asghari et al. (2015) [20]	Hamta4	English	Persian	Artificial
Ferrero et al. (2016) [18]	-	English	French Spanish	Artificial Simulated
Muneer, et al. (2019) [19]	CLEU	English	Urdu	Real
Asghari, et al. (2019)	Hamta-CL	English	Persian	Artificial Simulated
Haneef, et al. (2019) [21]	CLPD-UE-19	English	Urdu	Artificial Simulated
Mohtaj, et al. (2022) [22]	Hamta3	English	Persian	Artificial
Aljuaid, H. [32]	-	English	Arabic	Simulated
Stegmüller, J., et al. [33]	-	English	Spanish Japanese French Chinese	Real Simulated
Agarwal, B., et al. [34]	-	English	Hindi	Simulated
Bakhteev, O., et al. [35]	-	English Italian Swedish	Russian German Czech	Artificial Simulated
Zubarev, D., et al [36]	-	English	Russian	Artificial Simulated

[a] Language of source documents
[b] Language of suspicious documents
[c] Types of obfuscation

On the other hand, the artificial strategy for creating plagiarism detection corpora relies on employing some automatic synthesis to obfuscate original passages and generate paraphrased suspicious fragments. Some of the common operations to generate artificially paraphrased fragments include deletion of some words, shuffling words in sentences and addition of new words. Although most of the mono-lingual PD corpora have used simulated obfuscation methods to compile more realistic dataset, less effort have been made to generate simulated cases of plagiarism in a cross-lingual dataset. In this paper, we introduce a wide range of simulated and artificial obfuscation methods to create bi-lingual plagiarized fragments. In order to accurately evaluate plagiarism detection system, it should face various types of paraphrasing with different levels of complexity. The proposed strategies for obfuscating passages are listed below:

Simple Translation: In simple translation obfuscation, we combine topically related sentences extracted from a parallel corpus to compile plagiarized passage.

TABLE II. A SAMPLE FRAGMENT FROM EACH PROPOSED OBFUSCATION METHODS.

Type of obfuscation	English Passage	Persian Passage	Obfuscated passage
Simple translation (STR)	Susan Brown is a famous writer. She met her husband, Mr Johnson, in 1943 and she married him in 1944.	سوزان براون نویسنده‌ای مشهور است. او با شوهرش ، آقای جانسون ، در سال ۱۹۴۳ آشنا شد و در سال ۱۹۴۴ با او ازدواج کرد.	سوزان براون نویسنده‌ای مشهور است. او با شوهرش ، آقای جانسون ، در سال ۱۹۴۳ آشنا شد و در سال ۱۹۴۴ با او ازدواج کرد.
Artificial (ART)	When I was a student, I read a book whose title was Gone with the Wind, written by Margaret Mitchell.	وقتی دانشجو بودم کتابی خواندم که عنوانش برباد رفته نوشته مارگارت میشل بود.	زمانی بودم کتاب مطالعه کردم مارگارت میشل عنوان برباد رفته.
Paraphrasing (PAR)	Not only does an incompetent manager fail to run a company in excellent fashion, but he or she is unable to make a friendly relationship with their staff as well.	یک مدیر نالایق نه تنها نمی تواند یک شرکت را خوب اداره کند بلکه نمی تواند با کارمندانش به خوبی ارتباط برقرار کند.	مدیری که نمی تواند یک شرکت را خوب اداره کند و با کارمندانش به خوبی ارتباط برقرار کند، نالایق است.
Summarization (SUM)	Roger argued for the repeal of compulsory elementary and secondary school attendance laws with six arguments.. These are divided into three groups: First, education is for those who want to learn, and when including those who do not want to learn, everyone suffers. Second, grades will reflect effort and elementary teachers will not feel pressured to pass failing students. Third, schools would save money and face by eliminating compulsory attendance laws.	راجر موضوع رها شدن از قوانین حضور اجباری در مدارس ابتدایی و راهنمایی را در ۶ استدلال مشخص میکند. این استدلال‌ها در سه گروه قرار میگیرند. اول اینکه تحصیل برای کسانی است که می‌خواهند یاد بگیرند و با شامل شدن انهایی که نمی خواهند یاد بگیرند، همه رنج می‌برند. دوم اینکه، نمرات بازتابی از تلاش می‌باشد و معلمان مدارس ابتدایی احساس اجبار نمی کنند که دانش اموزان رد شده را قبول شده اعلام کنند. سوم اینکه، این مدارس با حذف قوانین حضور اجباری پول و شهرت خود را حفظ میکنند.	راجر موضوع رها شدن از قوانین حضور اجباری در مدارس ابتدایی و راهنمایی را در ۶ استدلال مشخص میکند که در سه گروه قرار میگیرند: عدم رضایت دانش اموزان به دلیل حضور افرادی که تمایلی به یادگیری ندارند و عدم اجبار معلمان برای پذیرفتن دانش اموزان مردود شده و حفظ پول و شهرت مدارس.
Pivot Translation (PIV)	With more than one million servers and data centers around the world, Google is able to process more than 1 billion search queries daily and its search engine is the most visited website globally as shown by the ranking international web.	با بیش از یک میلیون سرور و مراکز داده در حال حاضر در سراسر جهان، گوگل قادر به پردازش بیش از ۱۰۰۰ میلیون درخواست روزانه جستجو است و موتور جستجویش با بیشترین بازدید وب سایت در سراسر جهان در رتبه بندی وب بین المللی نشان داده شده است.	با بیش از یک میلیون سرور و مراکز داده در سراسر جهان، گوگل می‌تواند بیش از ۱ میلیارد درخواست جستجو در روز پردازش و موتور جستجوی خود بیشترین بازدید وب سایت-ها در جهان است، به عنوان نشان داده شده است رتبه بندی وب بین المللی است.
Splitting (SPL)	Widely reported, if somewhat doubtful, accounts from figures such as the famous Venetian traveler Marco Polo of the Chinese's willingness to trade with Europeans and the vast wealth that could result being through such contact makes the idea irresistible.	گزارشات گسترده توسط افراد مشهور مثل مارکوپولو، از اشتیاق مردم چین در تجارت با اروپایی‌ها و ثروتی که از این طریق بدست آمده است، که گاهاً نیز مورد قبول نبوده است، این ایده را غیرقابل انکار کرده است.	گزارشات گسترده‌ای توسط افراد مشهوری مثل مارکوپولو ارایه شده است. این گزارشات اگر چه گاهی مورد قبول نبوده است اما حاکی از اشتیاق مردم چین در تجارت با اروپایی‌ها و ثروتی که از این طریق بدست آمده است، می-باشد. درنتیجه این گزارشات، این ایده غیرقابل انکار خواهد بود.
Merging (MRG)	I have to support my family. I should find a job with high salary.	من مجبورم از خانواده ام حمایت کنم. من باید یک کار با حقوق بالا پیدا کنم.	من باید یک کار با حقوق بالا برای حمایت از خانواده ام پیدا کنم.

Artificial: In artificial obfuscation, we combine topically related sentences extracted from a parallel corpus. Then an artificial obfuscation in the target language is performed to create final plagiarized passages.

Paraphrasing: In paraphrasing obfuscation, at first we create plagiarized passages by combining topically related sentences from a parallel corpus. Then a human aided paraphrasing in the target language is performed. In other words, in this type of simulated obfuscation, a monolingual paraphrasing is performed in the target language.

Summarization: In summarization obfuscation, at the first step the English passage is translated, and then summarization in the target language is performed.

Pivot Translation: In Pivot Translation obfuscation, at first a translation from a source language L1 (i.e. English) to a different language L3 is performed, then it is translated back into the target language L2 (i.e. Persian).

Split: In Split obfuscation, after translating the passage to target language (Persian), the resulted sentence is divided into two or more sentences in the target language.

Merge: In Merge obfuscation, after translating the passage to the target language (Persian), the two or more resulting sentences are combined into one sentence.

The simple translation type of obfuscation consists of combining topically related sentences from a parallel corpus to generate both Persian and English passages. Since no additional changes have been applied on generated passages, this type of passage generation can be considered as "No Obfuscation" paraphrasing strategy. The goal is to generate simulated cases of literal translation from source language to target language for creating plagiarized passages.

Like simple translation type of obfuscation, the artificial type consists of generating passages by combining sentences from a parallel corpus. For creating more complicated cases of plagiarism, some automatic

synthesizing has been applied to Persian text to alter the passage. Some of the operations that are used to automatically paraphrase the Persian passages are the addition of new words, deletion of words, changing some words with synonyms, and shuffling words. FarsNet Persian semantic network [23] has been used to replace some words of the source fragment with their synonyms. This semantic network contains about 30,000 entries which are organized in about 20 thousand synsets. Most of the synsets in this semantic network are mapped to synsets in Princeton WordNet. Parsivar Persian text processing toolkit [24] has been used to perform pre-processing tasks such as tokenization, stemming, and POS-tagging. According to Mohtaj et al., [24] and based on their test conditions, the overall precision of tokenizer, POS-tagger and stemmer tools in Parsivar are 91%, 95% and 90% respectively.

The paraphrasing type of obfuscation consists of involving crowd-workers to rewrite Persian passages and generate fragments with the same meaning and different wording. The purpose is to simulate cases of paraphrasing when the author tries to conceal plagiarism by changing the wordings in the target language.

The summarization type of obfuscation tries to simulate cases in which the plagiarized passage from the source language would be shortened to make it difficult to detect. To this end, crowd-workers have been asked to summarize the Persian passage, keeping up the main concept of fragments. Unlike other types of obfuscation, the plagiarized passages for summarization type of paraphrasing have been generated from Persian and English "abstract section" of academic papers. The reason is to obtain more coherent and longer passages which are better candidates for summarization.

The pivot translation type of obfuscation includes translating passages in the source language (L1) to other languages (may be two or more other languages), and translate it back to the target language (L2) using machine translation APIs. Due to different language models in different languages, the pivot translation of passages can lead to paraphrasing them and creating fragments with different wording. Although this approach has been widely used for compiling mono-lingual plagiarism detection corpora, this is the first time that this method has been used for generating cross-lingual cases of paraphrasing. The merge and split types of obfuscation consist of applying syntactical changes to the target language (i.e. Persian) to paraphrase the passages.

The merge type of obfuscation includes combining two or more Persian sentences into one sentence. Moreover, the split type of obfuscation includes breaking one sentence into two or more sentences in Persian. These types of obfuscation are challenging, and can measure the performance of algorithms in the cases of choosing appropriate levels of granularity. Regarding the split of the sentences, for keeping the sentences to be syntactically correct, the crowd-workers were allowed to add conjunction words.

All of the mentioned fragments with different types of obfuscations have been created by choosing topically related sentences and combining them into one connected passage. In the next section, we will describe the steps for constructing the proposed corpus along with the statistical information of generated fragments.

IV. Corpus Construction

In this section, the method for compiling the proposed cross-lingual PD corpus is described in detail. Regarding the lack of variety of obfuscation in existing bilingual PD corpora, we introduce different types of obfuscation to simulate paraphrasing of text passage during scientific theft. We aim to construct a corpus that could be used for the evaluation of PD systems that reflects the re-use of scientific passages from a source language after translating and paraphrasing the passage to make it hard to find. Moreover, the resources for constructing and the statistics of compiled corpus are presented in this section.

1-1- Document Resources

The resources to construct the PD corpus play an important role to simulate a realistic corpus. Since the goal is to compile a cross-lingual corpus, multi-lingual resources should be used. Moreover, it is also intended to use open access resources to be able to share the corpus with the research community. We used Wikipedia articles for the main body of the corpus, since it includes all of the mentioned features. A small collection of scientific papers has also been added into the resources, so the corpus has a better coverage of various topics in scientific articles. Shortly, for constructing the corpus we have extracted raw text data from the following data resources:

- Topic-aligned Wikipedia pages (articles with English content along with their corresponding Persian content)
- The abstract part of Persian scientific papers and their equivalent English abstracts
- An English-Persian parallel corpus (Mizan [4])

It is worth mentioning that the first resource has been used to construct all of the source and suspicious documents and the last two ones have been exploited for constructing the plagiarized fragments. Among the selected document resources, the English passages are chosen as source documents and the Persian ones are selected as the suspicious ones.

To make the plagiarism cases more realistic and to make the detection process to be harder, cross-lingual topic modeling has been applied on source and suspicious documents. It leads to choosing the source document and the respective suspicious document from the same topic. The polylingual topic model (PLTM) has been used to cluster the documents in English and Persian into different categories [25]. This model is an extension of

latent Dirichlet allocation (LDA) to model the polylingual document tuples [26]. Each tuple includes a set of documents that are loosely equivalent to each other, but have been written in different languages, for example, corresponding Wikipedia articles in Persian and English. PLTM presumes that the documents in a tuple share the same tuple-specific distribution over topics. This is unlike LDA, in which each document is assumed to have its own document-specific distribution over topics. All of the documents (in both languages) have been split into 10 different topics.

To generate cases of cross-lingual plagiarized passages, topically related sentences from the selected parallel corpus have been combined into one passage. Since in real cases, plagiarism can occur in different lengths, so a wide range of case lengths should be considered to create plagiarized fragments. To simulate various distribution of lengths of plagiarism cases, the lengths of fragments are distributed between 20 and 300 words.

1-2- Fragment Generation and Obfuscation

As mentioned in previous sections, there are seven types of paraphrasing that have been proposed for obfuscating the passages named as Simple Translation, Artificial, Paraphrasing, Summarization, Pivot Translation, Split and Merge. The types of paraphrasing and a sample fragment from each proposed obfuscation method is shown in Table 2. For constructing the passages with the above-mentioned obfuscation types, an English-Persian parallel corpus and a number of Persian and English abstract of journal papers have been collected. It should be mentioned that the abstract part of journal papers are just used to generate summarization type of obfuscation. According to Table 3, it covers 8% of total fragments. Moreover, the ratio of number of fragments in each type of obfuscation with respect to the total number of fragments is selected as in [20]. A total number of 20 crowd workers from different ages and different level of educations have been employed to simulate some types of plagiarism. The demographic information of participants is presented in Table 4.

TABLE III. RATIO OF DIFFERENT TYPES OF OBFUSCATION IN THE PROPOSED CORPUS [20]

Obfuscation	No. of fragments in corpus	% of fragments in corpus
Simple Translation (STR)	498	29%
Artificial (ART)	495	29%
Paraphrasing (PAR)	185	10%
Summarization (SUM)	134	8%
Pivot Translation (PIV)	187	10%
Split (SPL)	144	6%
Merge (MRG)	58	5%

TABLE IV. DEMOGRAPHIC INFORMATION OF PARTICIPANTS

Age	20 – 30	55%
	30 – 40	35%
	40 – 50	10%
Gender	Male	65%
	Female	35%
Education	College	10%
	Bsc.	15%
	Msc.	60%
	PhD.	15%

Figure 1 shows the length difference of paraphrased fragments with respect to the original fragments for different types of paraphrasing. The left part of the graph shows the frequency of paraphrased fragments with lengths shorter than the original passages. As presented in the figure, summarization, merge and pivot translation types of obfuscation tend to have shorter length than the original ones. On the other hand, the suspicious fragments in split type of obfuscation are longer than their original ones in English.

The formula for computing the length ratio is demonstrated below:

$$LengthRatio = \frac{Len(S_{susp}) - Len(S_{src})}{Len(S_{src})} \quad (1)$$

In this equation, *Len* (S_{src}) and *Len* (S_{susp}) represent the number of characters in the original and paraphrased fragments, respectively.

1-3- Insertion of passages into documents

The last stage for compiling the CLPD corpus is to insert generated passages into related source and suspicious documents. In order to compile the corpus as realistic as possible, the obfuscated passages have been inserted into topically related documents. This can also result in preventing simple detection of plagiarized fragments by investigating topic alteration in source and suspicious documents. Besides considering the topic similarity, the ratio of plagiarized passages per document is a key factor to compile a more realistic corpus. To simulate different situations of real plagiarism, various ranges of plagiarism per document is considered to compile the proposed corpus. To this aim, we have considered a wide range of plagiarism ratios, from the low ratio of plagiarized fragments per suspicious document (hardly) to almost the entire content of document is plagiarism (entirely).

Figure 2 shows the position of generated plagiarized fragments in related source and suspicious documents. As depicted in the graph, the plagiarized fragments have been inserted into totally random positions in the documents.

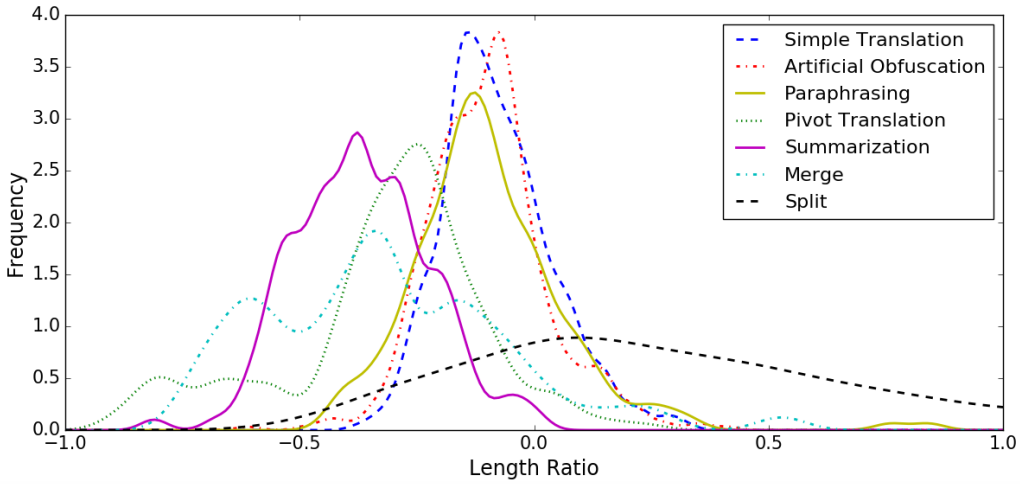

Figure 1. Distribution of length difference between original and paraphrased fragments

In the last step, the metadata information of constructed plagiarism cases is inserted into a specified file with XML data format. The resulted corpus can be effectively incorporated into an evaluation framework to investigate the performance of plagiarism detection algorithms. For each pair of suspicious and source documents, the corresponding XML file consists of the name of source and suspicious documents along with the following data for each plagiarism fragment:

- Type of obfuscation (none, random, simulated, pivot translation, summarization, merge, split or without plagiarism)
- Offset of the plagiarized fragment in suspicious document (offset of (S_{susp}))
- Length of the plagiarized fragment in suspicious document (length of (S_{susp}))
- Offset of the original fragment in the source document (offset of (S_{src}))
- Length of the original fragment in the source document (length of (S_{src}))

This XML meta-data can be used for evaluating the CLPD algorithms that use the proposed corpus for measuring their performance on plagiarism detection.

1-4- Corpus Evaluation

The compiled corpus should be analyzed from various aspects to be reliable enough to evaluate the plagiarism detection algorithms. There are some researches for validation and evaluation of PD corpora such as [27].

In order to evaluate a plagiarism detection corpus, in the first step they should be validated. Validation process is to investigate the correctness of the offset of plagiarism cases and also their length in comparison to the values in metadata XML files. Moreover, the relation between plagiarism cases in suspicious documents to the peer cases in the source documents is also randomly checked to ensure the correct their relatedness. The other validation task is to search for the random distribution of plagiarism cases across suspicious documents.

The second step is to evaluate the quality of the corpus. In the evaluation process, the quality of source documents and the suspicious documents along with the quality of plagiarism cases are investigated. There are four ways for the qualification of the corpus:

- Manual investigation: In this method, we manually investigate the qualities of some randomly selected plagiarism cases that have been generated using artificial obfuscation. Moreover, all of the simulated plagiarism cases are inspected to assure their quality.
- Extrinsic evaluation: The second method for corpus evaluation is to inspect various parts of the corpus by a basic plagiarism detection algorithm. So, we expect that the results of the algorithm show a lower accuracy for the highly obfuscated plagiarism cases.
- Evaluation through comparison with a standard corpus: In research investigated by [2], 10 basic n-gram similarity detection methods (n varies between 1 to 10) measure the similarity between source passages and corresponding plagiarized passages. Similarity measurement is accomplished by a simple VSM model using cosine similarity measure as follow:

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

Where $S_{src}.S_{susp}$ is the intersection between vectors of the original and paraphrased passages, and $\| S_{src}\|$ and and $\| S_{susp}\|$ are norms of original and paraphrased passages, respectively.

- Intrinsic evaluation using Pearson correlation coefficient: In this method, human judgment is used as an evaluation criterion. In a research investigated by [28], they have investigated the correlation between similarity measures and respective human judgments by two assessors. They have reported the

correlation across the languages. The results of this work also provide insights into the measurement of cross language similarity. The correlation between human judgments and computer assessments are extracted using Pearson correlation coefficient.

In this research we investigate the quality of the proposed corpus by checking generated fragments manually by the help of crowd-workers. To this end, all of the manually plagiarized fragments (e.g. Paraphrasing, Summarization, Split and Merge) have been checked manually by at least 3 different crowds to prevent probable errors. Because of low quality of generated texts, about 12 percent of manually paraphrased fragments have been removed by at least 2 reviews (in a majority voting basis) and have not been inserted into final source and suspicious documents. Moreover, 10 percent of automatically generated fragments (e.g. Artificial and Pivot Translation) have been chosen randomly to be checked manually.

In addition to investigating the quality of generated plagiarized fragments, the constructed corpus has been manually checked for the accuracy of offsets and annotations points of view. For this purpose, the position of some randomly selected plagiarized fragments in related source and suspicious documents and the annotated offset in the XML files have been manually investigated.

In the next section, we investigate the quality of our CLPD corpus by applying some extrinsic evaluation tasks.

V. EXPERIMENTS

In this section the performance of cross-lingual plagiarism detection algorithms is evaluated against separate sub-corpora containing different types of paraphrasing. Our goal is to assess how different levels of obfuscation complexity affect the performance of methods for detecting cases of plagiarism.

To investigate the effect of proposed corpus on the performance of plagiarism detection methods, various CLPD algorithms are applied on the corpus. The goal is to measure the impact of obfuscation complexity on the performance of PD methods. For this purpose, four different algorithms have been selected and applied on the corpus. Potthast et al. [2] proposed a taxonomy of five classes of approaches for cross-language text reuse detection which include Syntax-Based, Dictionary-Based, Parallel Corpora-Based, Comparable Corpora-Based and MT-Based models. We selected four state-of-the-art algorithms from the above-mentioned methods including CL-ESA, CL-KGA, CL-LSA, and T+MA to be applied on the proposed corpus. It should be mentioned that the syntax-based methods are based on lexical similarity between languages with similar syntactic structures (e.g., related European language pairs). However, due to lexical differences and different alphabets between distant languages such as English and Persian, these methods cannot be applied for detecting cases of similarity.

The usual measures for evaluating the performance of NLP algorithms are precision, recall and F-measure. In plagiarism detection, a character-level precision and recall is used. Besides these measures, another performance measure that characterizes the performance of a detection algorithm have been defined in [2] which determines whether a plagiarism passage is detected as a whole or has been detected in several pieces. Granularity quantifies whether the contiguity between plagiarized passages is properly recognized. To formulate this characteristic, the granularity of R under S is introduced as follow:

$$gran\ (S,R\ ) = \frac{1}{|S_R|}\sum_{s\in S_R}|R_S| \qquad (3)$$

In the above equation, the range of gran (S, R) is between [1, R], with 1 indicates the desired one-to-one correspondence and R indicates the worst case. The granularity, character-level precision, and character-level recall can be combined to an overall single score as the following equation:

$$Plagdet\ (S,R\ ) = \frac{F_1}{1+gran(S,R)} \qquad (4)$$

Where S is the set of plagiarism cases in suspicious documents, R denote the set of plagiarism that has been detected, and F1 denotes the F-Measure. This measure has been used to evaluate the performance of the mentioned methods against different parts of the proposed corpus.

A detailed description about the chosen algorithms and the performance results of the methods for detecting cases of plagiarism are presented as follow.

CL-ESA: Potthast et al. in [31] proposed cross-lingual explicit semantic analysis (CL-ESA) as a cross-lingual retrieval model. In this model, a document d_1 in language L_1 can be represented as an ESA vector d'_1, using cosine similarity with the index collection D_1 in the corresponding language L_1. Moreover, a document d_2 in language L_2 can be presented as vector d'_2 by computing the cosine similarity of d_2 with index collection D_2 in language L_2. The similarity between two documents under ESA model is defined as similarity between the resulted vectors. A cosine similarity measure is usually used for measuring the similarity. It should be mentioned that the cross-lingual topic-aligned documents should be used to compute the similarity across languages.

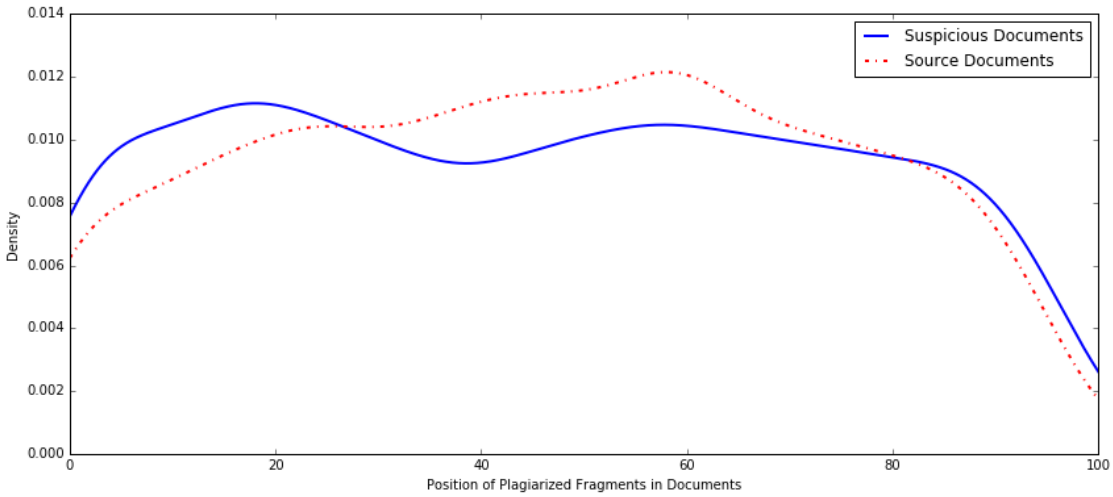

Figure 2. Relative Position of Plagiarized fragments in source and suspicions documents

As mentioned in [31], the document collection D should contain documents from a broad range of domains and also should have reasonable length. Since some of the Wikipedia documents fulfills both above-mentioned properties, so we used a collection of 20000 comparable Wikipedia articles. The chosen articles contain a wide range of topics and cover both Persian and English pages with lengths of more than 500 words.

In our experiments, we split the source documents and the corresponding suspicious documents into sentences. The sentences in the source and corresponding suspicious documents were embedded into vectors using CL-ESA algorithm. For this purpose, each sentence was compared with a collection of 20k documents by calculating cosine similarity measure. The English sentences were compared with English pages and the corresponding Persian sentences were compared against equivalents Persian pages. For detecting cases of plagiarism between documents, the cosine similarity between derived vectors in two documents is calculated using the cosine similarity measure. In Figure 3, we have represented the performance results of CL-ESA versus different types of paraphrasing in the proposed corpus.

CL-LSA: The objective of cross-lingual latent semantic Analysis (CL-LSA) is to construct a multilingual semantic space [29]. In our proposed CL-LSA method, we use translated documents (even manual or automatic translation) to construct a set of bi-lingual training document pairs. Since the training documents contain terms from both languages, so the resulting LSA model is defined under a bi-lingual vector space.

In our experiment based on CL-LSA, we have used a sentence aligned parallel corpus to train the model. Both documents were split into sentences for detecting cases of text similarity between source and suspicious documents. Each sentence has been converted into a low-dimensional LSA space by the constructed LSA model. The resulted vectors of sentences can be compared using cosine similarity measure to detect cases of similarity between source and suspicious documents. Figure 4 represents the performance results of CL-LSA versus different types of paraphrasing (different sub-corpora) in the proposed corpus.

CL-KGA: The goal of cross language knowledge graphs analysis is to exploit explicit semantics for documents representation [30]. This model provides a context model by creating knowledge graphs that expand and relate the initial concepts of the suspicious and source passages. Finally, in the generated semantic graph space, the similarity is calculated.

Given a source document and a suspicious document we compare document fragments based on the following steps as described in [30]:

We divide the original documents into set of fragments, using a 5-sentence sliding window with a 2-sentence step on the input documents. The passages are then lemmatized and labeled according to their grammatical category. In the next step, the knowledge graph is built from the labeled fragments using the BabelNet. Finally, we compare these graphs to measure the similarity between different fragments. Figure 5 shows the performance results of CL-KGA versus different types of obfuscation in the proposed corpus.

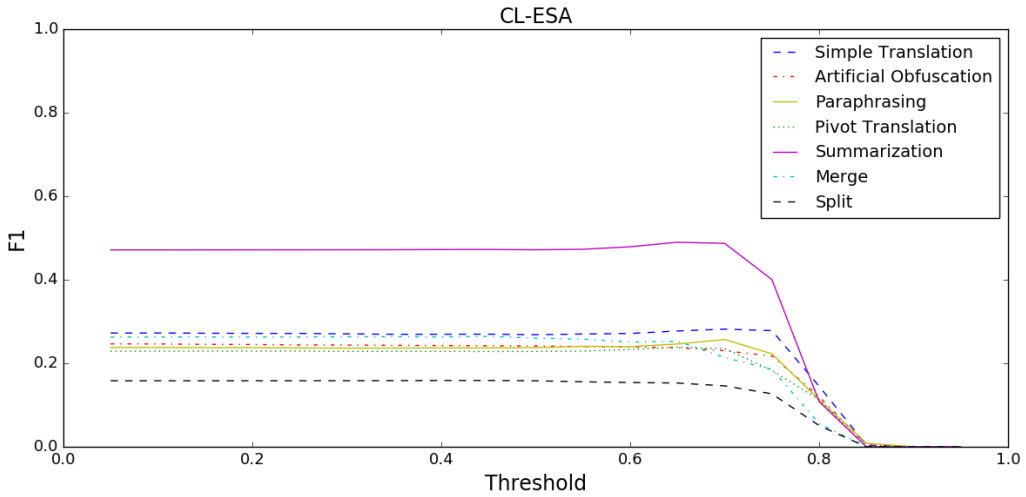

Figure 3. The performance results of CL-ESA in different sub-corpora

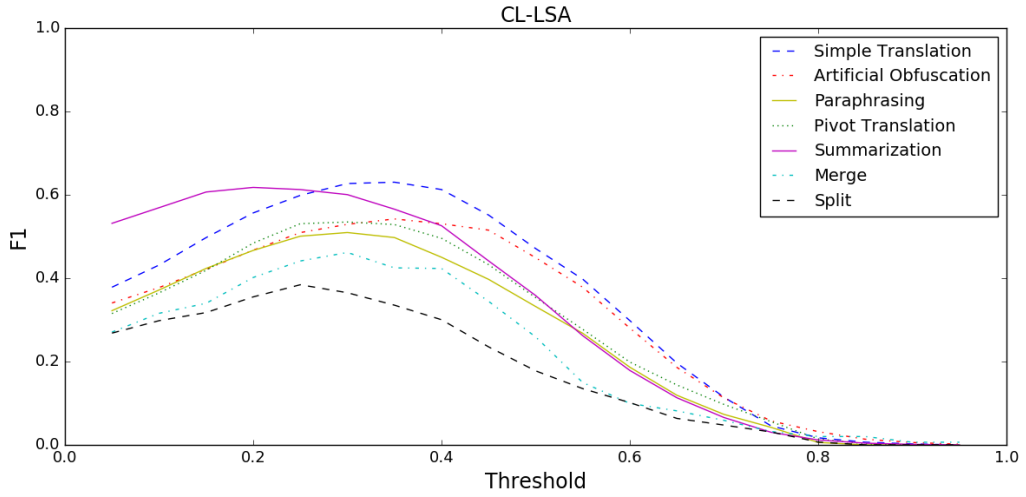

Figure 4. The performance results of CL-LSA in different sub-corpora.

T+MA: In the Translation plus Monolingual Analysis method, the suspicions documents were translated from Persian to English using the Google translate API. The Vector Space Model (VSM) algorithm is exploited to convert sentences from the acquired English documents and the source documents into vectors. Like previous models, the resulting sentence vectors are compared using a cosine similarity measure to detect similarity between the source and suspicious documents. Figure 6 demonstrates the performance results of T+MA versus different types of obfuscation in the proposed corpus.

In a general view, the performance of the algorithms is expected to be decreased while the obfuscation complexity will increase. In simple translation (STR), since there is little obfuscation in the passages, most of the algorithms achieve their best results in this type of obfuscation. The artificial obfuscation passages (ART) are constructed automatically, while the paraphrase obfuscation passages (PAR) are manually edited by the humans. Therefore, the PAR fragments are more complicated to find than ART fragments. As shown in the figures, the performance of the algorithms in artificial obfuscation (ART) are better with respect to paraphrase obfuscation (PAR). The process that is done on compiling Merge (MRG) and split (SPL) obfuscation, causes the structure of sentences to be messed up, whereas all of the algorithms work on a sequence of specific individual sentences to detect cases of plagiarism. So, the worst performance of the algorithms arises in these obfuscation types with respect to the other types of obfuscation. Furthermore, the performance of all of algorithms under MRG obfuscation is higher than SPL.

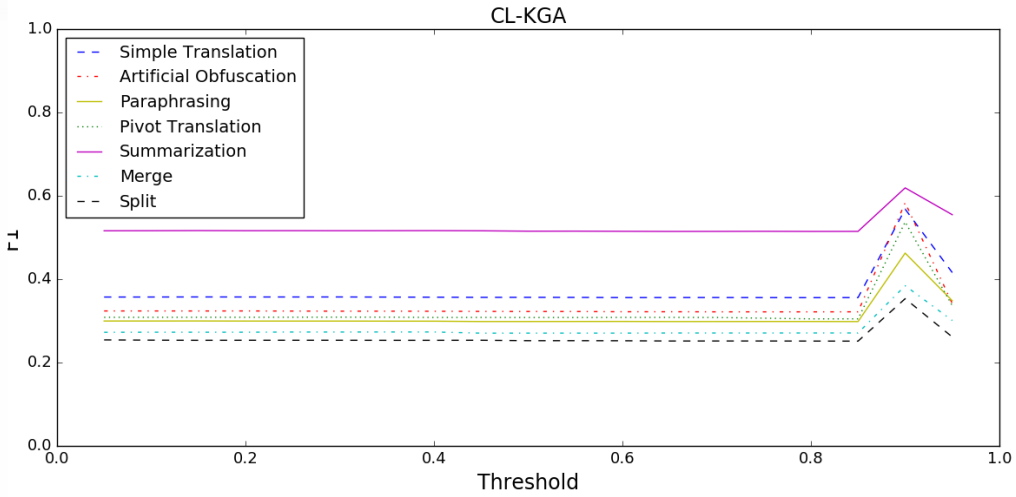

Figure 5. The performance results of CL-KGA in different sub-corpora

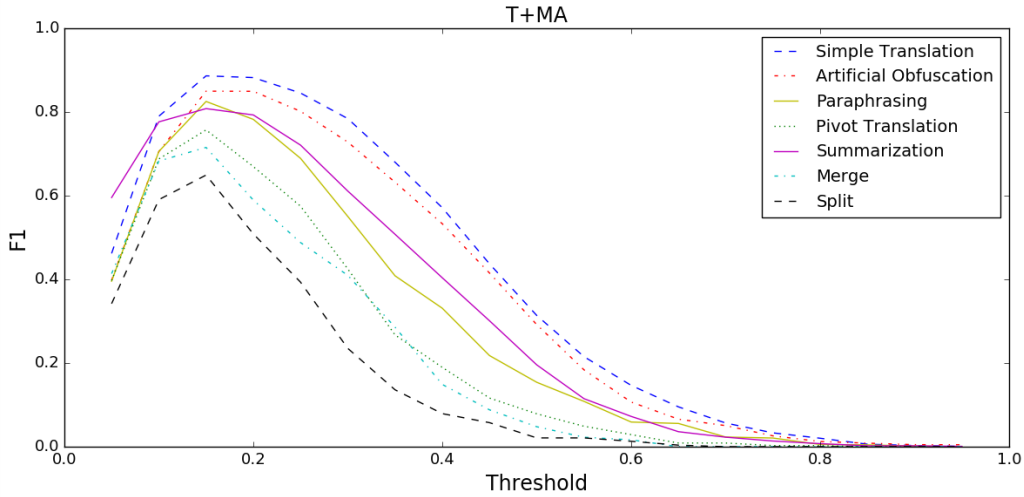

Figure 6. The performance results of T+MA in different sub-corpora

Although it seems that the most complex obfuscation is summarization, but since the summarized passages are relatively long while comparing to the MRG and SPL passages, so the performance of the algorithms in SUM reached better results with respect to merge and split obfuscation.

VI. Conclusion and Future Works

In this paper, we proposed a cross-lingual plagiarism detection corpus with seven different types of paraphrasing including Simple Translation (STR), Artificial (ART), Summarization (SUM), Paraphrasing (PAR), Pivot Translation (PIV), Split (SPL) and Merge (MRG). The proposed paraphrasing methods cover both artificial and simulated types of obfuscation. Moreover, cross-lingual topic modeling of documents has been used to simulate more realistic cases of plagiarism. The plagiarized fragments are inserted into topically related source and suspicious documents.

In order to evaluate our proposed corpus, we have used an extrinsic evaluation approach. For this purpose, we applied four state-of-the-art approaches in cross-lingual plagiarism detection (including CL-ESA, CL-KGA, CL-LSA and T+MA) on different parts of the corpus and the whole corpus as well. The results show that the performance of the algorithms decreases by increasing the obfuscation complexity. This corpus can be incorporated into an evaluation framework to study the performance of the CLPD algorithms. The corpus is freely available on the web for the research community.

In our future research we intend to add levels of obfuscation into this corpus in order to construct a multi-type, multi-level paraphrasing CLPD corpus.

Acknowledgement

The authors would like to thank the members of ITBM research group, ICT research Institute, ACECR. A special gratitude goes out to all the participants who helped us as crowd workers to construct simulated cases of plagiarism. Special thanks go to Dr Heshaam Faili for his valuable help along the way which greatly assisted this research.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Kashfi, Omid. 2018. MIZAN: A Large Persian-English Parallel Corpus. CoRR .Mimno D.,Wallach H., Naradowsky

[5] Ceska, Z., Toman, M., and Jezek, K., Multilingual Plagiarism Detection. Artificial Intelligence: Methodology, Systems, and Applications pp. 83–92 (2008).

[6] 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.

[7] 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.

[8] Potthast, Martin, Andreas Eiselt, Luis Alberto Barrón Cedeño, Benno Stein, and Paolo Rosso. "Overview of the 3rd international competition on plagiarism detection." In CEUR workshop proceedings, vol. 1177. CEUR Workshop Proceedings, 2011.

[9] Pinto, David, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, and Paolo Rosso. "A statistical approach to crosslingual natural language tasks." Journal of Algorithms 64, no. 1 (2009): 51-60.

[10] 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.

[11] Barrón-Cedeno, Alberto, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, and Mark Stevenson. "Pan@ fire: Overview of the cross-language !ndian text re-use detection competition." In Multilingual 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, December 2-4, 2011

[12] Alfikri, Z. F., and Purwarianti, A., The construction of Indonesian-English cross language plagiarism detection system using fingerprinting technique. Jurnal Ilmu Komputer dan Informasi 5: 16–23 (2012).

[13] Pataki, Máté. "A new approach for searching translated plagiarism." (2012): 49.

[14] Arefin, Mohammad Shamsul, Yasuhiko Morimoto, and Mohammad Amir Sharif. "BAENPD: A Bilingual Plagiarism Detector."J. Comput. 8, no. 5 (2013): 1145-1156.

[15] 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.

[16] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., & Potthast, M. (2018, February). Algorithms and Corpora for Persian Plagiarism Detection. In Text Processing: FIRE 2016 International Workshop, Kolkata, India, December 7–10, 2016, Revised Selected Papers (Vol. 10478, p. 61). Springer

[17] Hanif, I., Muhammad, R., Nawab, A., Arbab, A., Jamshed, H., Riaz, S., & Munir, E. U. (2015). Cross-Language Urdu-English (CLUE) Text Alignment Corpus Notebook for PAN at CLEF 2015. In CLEF (Notebook Papers/LABs/Workshops) Working Notes.

[18] Ferrero, J., Agnes, F., Besacier, L., & Schwab, D. (2016, May). A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4162-4169).

[19] Muneer, I., Sharjeel, M., Iqbal, M., Nawab, R. M. A., & Rayson, P. (2019). CLEU-A Cross-language english-urdu corpus and benchmark for text reuse experiments. Journal of the Association for Information Science and Technology, 70(7), 729–741.

[20] Asghari, Habibollah, Omid Fatemi, Salar Mohtaj, Heshaam Faili, and Paolo Rosso. "On the use of word embedding for cross language plagiarism detection."Intelligent Data Analysis 23, no. 3 (2019): pp. 661-680.

[21] Haneef, I., Nawab, A., Muhammad, R., Munir, E. U., & Bajwa, I. S. (2019). Design and Development of a Large Cross-Lingual Plagiarism Corpus for Urdu-English Language Pair. Scientific Programming, 2019.

[22] Mohtaj, S., & Asghari, H. (2022). A corpus for evaluation of cross language text re-use detection systems. Journal of Information Systems and Telecommunication (JIST), 3(39), 169.

[23] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., ... & Assi, S. M. (2010, January). Semi automatic development of farsnet; the persian wordnet. In Proceedings of 5th global WordNet conference, Mumbai, India (Vol. 29).

[24] Mohtaj, S., Roshanfekr, B., Zafarian,A., and Asghari. H., "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,

[25] Mimno, David, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. "Polylingual topic models." In Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 880-889. 2009.

[26] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of machine learning research 3, no. Jan (2003): 993-1022.

[27] Zarrabi, V., Rafiei, J., Khoshnava, K., Asghari, H., and Mohtaj, S. Evaluation of Text Reuse Corpora for Text Alignment Task of plagiarism Detection. Conference and Labs of the Evaluation forum (2015).

[28] 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.

[29] Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997, March). Automatic cross-language retrieval using latent semantic indexing. In AAAI spring symposium on cross-language text and speech retrieval (Vol. 15, p. 21). Stanford, CA, USA: Stanford University.

[30] Franco-Salvador, M., Gupta, P., & Rosso, P. (2014). Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. Bridging Between Information Retrieval and Databases: PROMISE Winter School 2013, Bressanone, Italy, February 4-8, 2013. Tutorial Lectures, 227-236.

[31] Potthast, Martin, Benno Stein, and Maik Anderka. "A wikipedia-based multilingual retrieval model." In European conference on information retrieval, pp. 522-530. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[32] Aljuaid, H. (2020). Cross-Language Plagiarism Detection using Word Embedding and Inverse Document Frequency

(IDF). International Journal of Advanced Computer Science and Applications, 11(2).

[33] Stegmüller, J., Bauer-Marquart, F., Meuschke, N., Ruas, T., Schubotz, M., & Gipp, B. (2021). Detecting Cross-Language Plagiarism using Open Knowledge Graphs. arXiv preprint arXiv:2111.09749.

[34] Agarwal, B., Gupta, M. K., Sharma, H., & Poonia, R. C. (2023). Siamese-Based Architecture for Cross-Lingual Plagiarism Detection in English–Hindi Language Pairs. Big Data, 11(1), 48-58.

[35] Bakhteev, O., Chekhovich, Y., Grabovoy, A., Gorbachev, G., Gorlenko, T., Grashchenkov, K., ... & Sakharova, A. (2023, January). Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works. In Proceedings from the European Conference on Academic Integrity and Plagiarism (pp. 143-161).

[36] Zubarev, D., Tikhomirov, I., & Sochenkov, I. (2021, October). Cross-lingual plagiarism detection method. In International Conference on Data Analytics and Management in Data Intensive Domains (pp. 207-222). Cham: Springer International Publishing.

Habibollah Asghari received his Ph.D. degree in computer engineering from department of electrical and computer engineering, school of engineering, University of Tehran, in 2019, with a focus on Text Similarity Detection and Natural Language Processing. He currently serves as an associate professor in department of Advanced Information Systems, ICT research Institute (ACECR). His research interests include Text Mining and Plagiarism Detection

Salar Mohtaj received his Ph.D. degree in Computer Science with a focus on Natural Language Processing in 2024. He is currently a Senior Researcher in the Speech & Language Technology group at the German Research Center for Artificial Intelligence (DFKI). His research interests include Transfer Learning, Harmful Content Detection, and the Evaluation of Large Language Models.