Hindawi
Scientific Programming
Volume 2019, Article ID 2962040, 11 pages
https://doi.org/10.1155/2019/2962040

Research Article

Design and Development of a Large Cross-Lingual Plagiarism Corpus for Urdu-English Language Pair

Israr Haneef,[1] Rao Muhammad Adeel Nawab,[2] Ehsan Ullah Munir,[1] and Imran Sarwar Bajwa[3]

[1]*Department of Computer Science, COMSATS Institute of Information Technology, Wah Campus, Wah Cantonment, Pakistan*
[2]*Department of Computer Science, COMSATS Institute of Information Technology, Lahore Campus, Lahore, Pakistan*
[3]*Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan*

Correspondence should be addressed to Imran Sarwar Bajwa; imran.sarwar@iub.edu.pk

Received 17 December 2018; Accepted 25 February 2019; Published 17 March 2019

Guest Editor: Vicente García-Díaz

Cross-lingual plagiarism occurs when the source (or original) text(s) is in one language and the plagiarized text is in another language. In recent years, cross-lingual plagiarism detection has attracted the attention of the research community because a large amount of digital text is easily accessible in many languages through online digital repositories and machine translation systems are readily available, making it easier to perform cross-lingual plagiarism and harder to detect it. To develop and evaluate cross-lingual plagiarism detection systems, standard evaluation resources are needed. The majority of earlier studies have developed cross-lingual plagiarism corpora for English and other European language pairs. However, for Urdu-English language pair, the problem of cross-lingual plagiarism detection has not been thoroughly explored although a large amount of digital text is readily available in Urdu and it is spoken in many countries of the world (particularly in Pakistan, India, and Bangladesh). To fulfill this gap, this paper presents a large benchmark cross-lingual corpus for Urdu-English language pair. The proposed corpus contains 2,395 source-suspicious document pairs (540 are automatic translation, 539 are artificially paraphrased, 508 are manually paraphrased, and 808 are nonplagiarized). Furthermore, our proposed corpus contains three types of cross-lingual examples including artificial (automatic translation and artificially paraphrased), simulated (manually paraphrased), and real (non-plagiarized), which have not been previously reported in the development of cross-lingual corpora. Detailed analysis of our proposed corpus was carried out using n-gram overlap and longest common subsequence approaches. Using Word unigrams, mean similarity scores of 1.00, 0.68, 0.52, and 0.22 were obtained for automatic translation, artificially paraphrased, manually paraphrased, and nonplagiarized documents, respectively. These results show that documents in the proposed corpus are created using different obfuscation techniques, which makes the dataset more realistic and challenging. We believe that the corpus developed in this study will help to foster research in an underresourced language of Urdu and will be useful in the development, comparison, and evaluation of cross-lingual plagiarism detection systems for Urdu-English language pair. Our proposed corpus is free and publicly available for research purposes.

1. Introduction

In cross-lingual plagiarism, a piece of text in one (or source) language is translated into another (or target) language by neither changing the semantics and content nor referring the origin [1, 2]. Cross-lingual plagiarism detection is a challenging research problem due to various reasons. Firstly, machine translation systems are available online free of cost such as Google Translator (https://translate.google.com/) to translate a document written in one language into another language. Secondly, the Web has become a hub of multi-lingual resources. For example, Wikipedia contains articles in more than 200 languages on same topics (http://en.wikipedia.org/wiki/wikipedia Last visited 10-02-2019). Thirdly, people might be often interested to write in another language which is different from their native language. Consequently, all these factors contribute to an environment, which makes it easier to commit cross-lingual plagiarism and difficult to detect it.

The task of plagiarism can be broadly categorized into two categories [3]: (1) intrinsic plagiarism analysis and (2) extrinsic plagiarism analysis. In the former case, a single document is examined to identify plagiarism in terms of variation of an author(s)'s writing style. The fragment(s) for text which is significantly different from other fragments in a document is a trigger of plagiarism. Mostly stylometric-based features are modeled to detect such plagiarism. In the latter case, we are provided with a document which is suspected to contain plagiarism (suspicious document) and source collection. The aim is to identify fragments of text(s) in the suspicious document which are plagiarized and their corresponding source fragments from the source collection. Extrinsic plagiarism can be further divided into (1) monolingual–both source and plagiarized texts are in the same language and (2) cross-lingual plagiarism–source and plagiarized texts are in different languages. In case of cross-lingual plagiarism, a source text can be translated either automatically or manually, and after translation, it can be either used verbatim or rewritten for plagiarism [4].

To develop and evaluate Cross-Lingual Plagiarism Detection (CLPD) methods, standard evaluation resources are needed. Majority of CLPD corpora are developed for English, European, and some other languages (http://www.webis.de/research/corpora-Last-visited-10-02-2019). In addition, none of the existing cross-lingual corpus contains a mix of artificial, simulated, and real examples, which is necessary to make a realistic and challenging corpus. The problem of CLPD has not been thoroughly explored for South Asian languages such as Urdu, which is a widely spoken by a large number of people around the globe. Urdu is the first language of about 175 million people around the world and particularly spoken in Pakistan, India, Bangladesh, South Africa, and Nepal (http://www.ethnologue.com/language/urd, last visited: 20-02-2019). It is written from right to left like Arabic script. Urdu language usually follows Nastalique writing style [5]. However, Urdu is an under-resourced language in terms of computational and evaluation resources.

The main objectives of this study are threefold: (1) to develop a large benchmark cross-lingual corpus for Urdu-English language pair, which contains a mix of artificial, simulated, and real examples, (2) to carry out linguistic analysis of the proposed corpus to get insights into the edit operations used in cross-lingual plagiarism, and (3) to carry out detailed empirical analysis of the proposed corpus using n-gram Overlap and Longest Common Subsequence approaches to investigate whether the documents in the corpus are created using different obfuscation techniques. There are total 2,398 source-suspicious document pairs in our proposed corpus. Source documents are in Urdu language, and suspicious ones are in English. The source-suspicious document pairs are categorized into two main categories: (1) plagiarized (1,588 document pairs) and (2) nonplagiarized (810 document pairs). The plagiarized documents are created using three obfuscation strategies: (1) automatic translation (540 document pairs), (2) artificial paraphrasing (540 document pairs), and (3) manual paraphrasing (508 document pairs). The documents in our proposed corpus are from various domains including Computer Science, Management Science, Electrical Engineering, Physics, Psychology, Countries, Pakistan Studies, General Topics, Zoology, and Biology, which makes the corpus more realistic and challenging. We also carried out linguistic and empirical analysis of our proposed corpus.

Our proposed corpus will be beneficial for (1) fostering and promoting research in a low resourced language—Urdu, (2) enabling us to make a direct comparison of existing and new CLPD methods for Urdu-English language pair, (3) developing and evaluate new methods for CLPD for Urdu-English language pairs, and (4) developing a bilingual Urdu-English dictionary using our proposed corpus. Furthermore, our proposed corpus is free and publicly available for research purposes.

The rest of this paper is organized as follows: Section 2 summarizes the related work on existing corpora for CLPD. Section 3 describes the corpus generation process, including source documents collection, levels of rewriting, creation of suspicious documents, and standardization of the corpus. Section 4 presents the linguistic analysis of our proposed corpus. Section 5 presents a deeper empirical analysis of the corpus. Finally, Section 6 concludes the paper.

2. Related Work

In the literature, efforts have been made to develop benchmark corpora for CLPD. One of the prominent efforts is the series of PAN (http://pan.webis.de/, last visited: 20-02-2019) (a forum of scientific events and shared tasks on digital text forensic) competitions. A number of frameworks for cross-lingual plagiarism evaluation are also proposed by researchers for this forum [6, 7]. The main outcome of these competitions is a set of benchmark corpora for mono- and cross-lingual plagiarism detection. The majority of plagiarism cases, in these corpora, are monolingual (90%), and remaining 10% are cross-lingual such as English-Persian and English-Arabic and other language pairs. Almost 80% of cross-lingual plagiarism cases, in these corpora, are generated using automatic translation, and the rest are generated using manual translation. PAN cross-lingual corpora have been developed for two language pairs: English-Spanish and English-German.

The relevant literature presents a number of benchmark CLPD corpora for languages like Indonesian-English [8], Arabic-English [9], Persian-English [10], and English-Hindi [11]. Developing such a resource for especially under-resourced languages is an active research area [12, 13]. Parallel corpora have also been developed and used in [14] for the automatic translation purpose in cross-lingual domain. CLPD systems based on these corpora and other approaches are also proposed in the literature [15]. Most of these approaches used syntax-based plagiarism detection methods, but at the same time, semantic-based plagiarism detection approaches were also applied for the purpose. Savador et al. used semantic plagiarism detection approach using the graph analysis method for cross-language plagiarism detection. It is a language-independent model for plagiarism detection applied to the Spanish-English and German-English domains [16].

Cross Language Indian Text Reuse (CLITR) task has been designed in conjunction with Forum for Information Retrieval Evaluation (FIRE) to detect cross-lingual plagiarism for English-Hindi language pair. The corpus is divided into training and test segments in which source documents are in English and suspicious documents are in Hindi.

The training and test collection both include 5032 source files in English while 198 suspicious files in training and 190 suspicious files are in Hindi (http://www.uni-weimar.de/medien/webis/events/panfire-11/panfire11-web/, last visited: 20-08-2018). Corpora have also been developed for performance evaluation of cross-language information retrieval (CLIR) systems [17], while Kishida [18] raised technical issues of this domain. Moreover, different plagiarism detection tasks like text alignment and source retrieval are designed based on these corpora's, and overview of these tasks are being consistently (yearly) been published by PAN@ CLEF forum [19, 20].

The JRC-Acquis Multilingual Parallel Corpus has been used by Potthast et al., to apply CLPD approaches. As many as 23,564 parallel documents are constructed in the corpus that is extracted from legal documents of European Union [21, 22]. Out of 22 languages in legal document collection, only 5 including French, Germen, Polish, Dutch, and Spanish were selected to generate source-suspicious document pair (English language was used as source language). Comparable Wikipedia Corpus is another dataset used for the evaluation of CLPD methods. The corpus contains 45,984 documents.

Benchmark cross-lingual corpora have been developed using two approaches: (1) automatic translation and (2) manual translation. PAN corpora are created using both approaches for English-Spanish and English-German language pairs. However, the majority of cross-lingual cases are generated using automatic translation, and only a small number of them are generated using manual translation.

CLITR Corpus is generated using both automatic and manual translations: Near copy/exact copy documents are created using automatic translation, whereas heavy revision (HR) documents are created using manual paraphrasing of automatic translations of source texts. Again, this corpus only contains 388 suspicious documents, and it is created for English-Hindi language pair.

Two cross-lingual corpora used in plagiarism detection task are (1) JRC-EU Corpus and (2) Fairy Tale Corpus [21, 22]. JRC-EU cross-lingual corpus consists of randomly extracted 400 documents from the legislation reports of European Union which includes 200 English source documents and 200 Czech documents. Fairy-tale corpus contains 54 documents: 27 in English and 27 in Czech. Ceska et al. also used these corpuses for CLPD task [23].

In a previous study, we developed a corpus for the PAN 2015 Text Alignment task (we named it CLUE Corpus) [24]. In that corpus, there are total 1000 documents (500 are source documents and 500 are suspicious documents). Among the suspicious collections, 270 documents are plagiarized using 90 source-plagiarized fragment pairs, while the remaining 230 suspicious documents are nonplagiarized. Note that this corpus contains simulated cases of plagiarism, which were inserted into suspicious document to generate plagiarized documents. The CLUE Corpus can be used for the development and evaluation of CLPD systems for English-Urdu language pair for the text alignment task only as described by PAN organizers.

To conclude, the relevant literature presents the majority of CLPD corpora for English and other European languages. Moreover, these are mainly created using comparable documents, parallel documents, and automatic translations, which are not realistic examples for cross-lingual plagiarism. This study contributes a large benchmark corpus (containing 2,398 source-suspicious document pairs) for CLPD in Urdu-English language domain. Note that the 270 fragment pairs used in the development of CLUE Corpus are also included in this corpus.

3. Corpus Generation

This section describes the process for construction of a benchmark corpus for CLPD for Urdu-English language pair (hereafter called CLPD-UE-19 Corpus) including collection of source texts, levels of rewrite used in creating suspicious documents, creation of suspicious documents, and standardization of corpus and corpus characteristics.

3.1. Collection of Source Texts. Urdu is an underresourced language as large repositories of digital texts in this language are not readily available for the research purposes. Urdu newspapers in Pakistan mostly publish news stories in images format which is not suitable for text processing. Therefore, to collect realistic, high-quality, and diversified source articles for generating CLPD-UE-19 Corpus, we selected Wikipedia[1] as a source. Wikipedia is a free and publicly available, multitopic, and multilingual resource. Initially, Wikipedia contains an article in multiple languages which makes it possible to be considered as a comparable corpus. AJ Head investigated the potential use of Wikipedia for course-related search by students [25]. Martinez also investigated the cases where Wikipedia is mainly used for copy and paste plagiarism cases [26]. Wikipedia articles are taken as source documents for generating cross-lingual plagiarism detection corpus for Hindi-English language pair [27].

Plagiarism is a serious problem, particularly in higher educational institutions [28–31]. Therefore, CLPD-UE-19 Corpus focuses on plagiarism cases generated by university students. Table 1 shows the domains from which Wikipedia (http://ur.wikipedia.org/wiki/urdu) source articles are collected to generate CLPD-UE-19 Corpus. Apart of it, 270 source-suspicious document pairs were used in the creation of the CLUE Corpus [24].

These domains include Computer Science, Management Science, Electrical Engineering, Physics, Psychology, Countries, Pakistan Studies, General Topics, Zoology, and Biology. As can be noted, these articles are on a wide range of topics, which makes the CLPD-UE-19 Corpus more realistic and challenging.

The amount of text reused for creating a plagiarized document can vary from a phrase, sentence, and paragraph to the entire document. It is also likely that to hide

Table 1: Domains from which Wikipedia source articles were selected in creating our proposed CLPD-UE-19 Corpus.

Domain	Major topics
Computer science	Free software, binary numbers, open source, database normalization, robotics, artificial intelligence, MSN, Google, Yahoo, WhatsApp, Android, Facebook, Twitter, RUBY language, daily motion, HTML, mobile apps, Gmail, Skype, and others
General topics	Globalization, muhammad iqbal, global warming, capitalism, mosque, bookselling, Pakistan air force, cricket, fashion, Lahore Fort, capitalism, Badshahi Masjid, and two-nation theory
Electrical engineering	Electricity, magnetism, and conducting materials
Management science	Trade and finance
Physics	Atoms and scientists
Psychology	Neurology, psycho diseases, and enlightenment
Countries	Politics and trade of different countries (mostly African)
Pakistan studies	History of Pakistan and Indo-Pak partition
Zoology	Animals, food, and living styles
Biology	Natural organisms, living cells, and DNA

plagiarism, a plagiarist may reuse the texts of different sizes from different sources. Therefore, the size of source documents is varied. The length of a source text may fall into one of the three categories: (1) small (1–50 words), (2) medium (50–100 words), and (3) large (100–200 words).

3.2. Levels of Rewrite. The proposed corpus contains two types of suspicious documents: (1) plagiarized and (2) nonplagiarized. The details of these are as follows.

3.2.1. Plagiarized Documents. A plagiarized document in CLPD-UE-19 Corpus falls into one of the three categories: (1) automatic translation, (2) artificially paraphrased copy, and (3) manually paraphrased copy. The reason for creating plagiarized documents with three different levels of rewrite is that a plagiarist is likely to use one of the three above-mentioned approaches for creating a plagiarized document using existing document(s) for cross-lingual settings.

(a) Automatic Translation. Using this approach, plagiarized documents (in English) are created by automatically translating the source texts (in Urdu) using Google Translator (https://translate.google.com/, last visited: 20-02-2019). Note that Google Translator has been effectively used in earlier research studies [32, 33].

(b) Artificially Paraphrased Copy. This approach aims to create artificially paraphrased cases of cross-lingual plagiarism in two steps. A source text (in Urdu) is translated automatically into English using Google Translator in the first step. After that, an automatic text rewriting tool is used to paraphrase the translated text, which results in an artificially paraphrased copy of the original text. For this study, we explore various free and publicly available text rewriting tools. Among the available tools, we found that two of them have the highest number of visitors per day: (1) Spinbot text rewriting tool (http://www.spinbot.net/) with an average number of 26 k visitors per day and (2) Article Rewriter text rewriting tool (http://articlerewritertool.com/) with an average number of 45 k visitors per day reported by Alexa (this is a ranking system set by alexa.com (a subsidiary of amazon.com) that basically audits and makes public the frequency of visits on various websites) as compared to other tools like http://paraphrasing-tool.com/, etc.

(c) Manually Paraphrased Copy. Using this approach, the plagiarized document were created by manually translating and paraphrasing the original texts.

3.2.2. Nonplagiarized. Wikipedia is a comparable corpus and contains an article in multiple languages. It is notable that these articles are not translations of each other. To generate nonplagiarized cases, similar fragments of texts were manually identified from English and Urdu Wikipedia articles on the same topic.

The assumption is that although English and Urdu Wikipedia articles are written on the same topic, they are independently written by two different authors. Therefore, similar fragments of English-Urdu texts can serve as independently written cross-lingual document pairs.

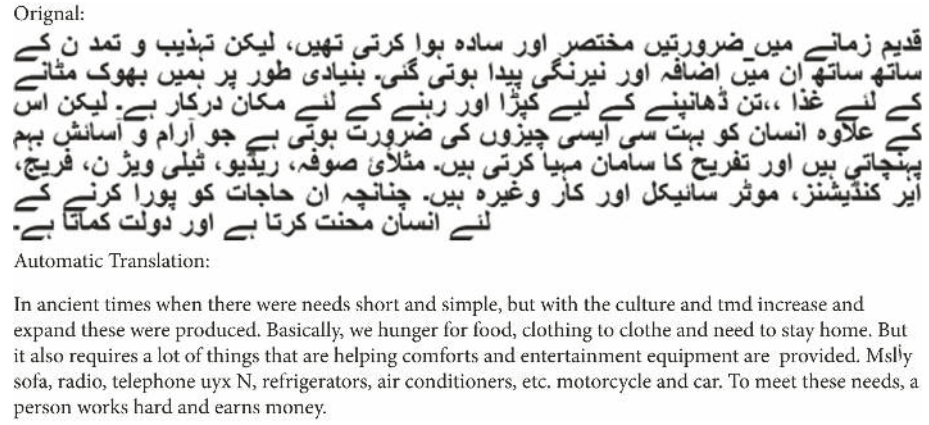
As far as we are aware, the proposed methods used for creating cross-lingual plagiarism cases of artificially paraphrased plagiarism and Nonplagiarism have not been previously used for creating cross-lingual plagiarism cases in any other language pair.

3.3. Generation of Suspicious Texts. Crowdsourcing is a process of performing a task in collaboration of a large number of people usually working as a remote user. It can be done with a group of people, small teams or even individuals. Generating a large benchmark CLPD corpus is not a trivial task. Therefore, we use the crowdsourcing approach to generate suspicious texts with four levels of rewriting. Examples of manually paraphrased copy and nonplagiarized are generated by participants (volunteers), who are graduate-level university students (masters and M Phil). All the participants are native speakers of Urdu. As the medium of instruction in university and colleges is English, students have a high level of proficiency in English language too.

The majority of the participants are from the English department, and hence are well aware of paraphrasing techniques.

However, for better quality, they were provided with examples of paraphrasing. The plagiarized documents generated by volunteers were manually examined, and low-quality documents were discarded.

3.4. Examples of Cross-Lingual Plagiarism Cases from CLPD-UE-19 Corpus. Figure 1 presents an example of source-


Orignal:

قدیم زمانے میں ضرورتیں مختصر اور سادہ ہوا کرتی تھیں، لیکن تہذیب و تمد ن کے ساتھ ساتھ ان میں اضافہ اور نیرنگی پیدا ہوتی گئی۔ بنیادی طور پر ہمیں بھوک مٹانے کے لئے غذا ،،تن ڈھانپنے کے لیے کپڑا اور رہنے کے لئے مکان درکار ہے۔ لیکن اس کے علاوہ انسان کو بہت سی ایسی چیزوں کی ضرورت ہوتی ہے جو آرام و آسائش بہم پہنچاتی ہیں اور تفریح کا سامان مہیا کرتی ہیں۔ مثلاً صوفہ، ریڈیو، ٹیلی ویژ ن، فریج، ایر کنڈیشنز، موٹر سائیکل اور کار وغیرہ ہیں۔ چنانچہ ان حاجات کو پورا کرنے کے لئے انسان محنت کرتا ہے اور دولت کماتا ہے۔

Automatic Translation:

In ancient times when there were needs short and simple, but with the culture and tmd increase and expand these were produced. Basically, we hunger for food, clothing to clothe and need to stay home. But it also requires a lot of things that are helping comforts and entertainment equipment are provided. Msly sofa, radio, telephone uyx N, refrigerators, air conditioners, etc. motorcycle and car. To meet these needs, a person works hard and earns money.

Figure 1: An example of plagiarized document created using automatic translation approach.

plagiarized document pair from CLPD-UE-19 Corpus created using automatic translation approach. As can be noted, the translated text is not an exact copy of the original one. The possible reason for this is that Urdu is an underresourced language, and machine translation systems for Urdu-English language pair are not matured compared to other language pairs. Consequently, the translated text seems to be a near copy of the original text instead of an exact copy. Moreover, it can also be observed from the translated document that for few words for which Google Translator does not find any equivalent word in English, it merely replaces the pronunciation of that word with English homonyms, for instance, is replaced with *tmd* and is replaced with *Msly*. To conclude, the overall quality of Google Translator seems to be good considering the complexity in translating Urdu text to English.

Figure 2 shows an example of plagiarism document where automatic translation of a source document is further altered by an automatic rewriting tool to get artificially plagiarized copy of the source document. It can be observed from this example that automatic text rewriting tool has replaced the words by appropriate synonyms (the words presented in Italics are synonyms of original words). However, the text rewriting tool does not alter the order of text. The alteration in the translated text is carried out by rewriting tool which further increases the level of rewriting and makes it difficult to identify similarity between source-plagiarized text pairs.

A sample plagiarized document generated using the manually paraphrased copy approach is shown in Figure 3, which is a very well paraphrased content. Different text rewriting operations have been applied by the participants to paraphrase the original text including synonym replacement, sentence merging/splitting, insertion/deletion of text, word reordering. Consequently, the source-plagiarized text pairs are semantically similar but different at surface level, which makes the CLPD task even more challenging.

A nonplagiarized source-suspicious document pair from the CLPD-UE-19 Corpus is shown in Figure 4. The text is topically related, but independently written. The inclusion of more introductory sentences and last sentence reflects that both texts are written in different contexts.

3.5. Corpus Characteristics. Table 2 presents the detailed statistics of the proposed corpus. In this table, AT, APC, MPC, and NP represent automatic translation, artificially paraphrased copy, manually paraphrased copy, and non-plagiarized, respectively. There are total 2,398 source-suspicious document pairs in the corpus, 810 are non-plagiarized and 1,588 are plagiarized. Among the plagiarized document pairs, 540 are automatically translated, 540 are artificially paraphrased, and 508 are manually paraphrased. Above statistics show that the corpus contains a large number of documents for both plagiarized and nonplagiarized cases. Also, the documents for four different levels of rewrite in the proposed corpus are almost balanced. The CLPD-UE-19 Corpus is standardized in XML format and publicly available for research purposes (the CLPD-UE-19 Corpus is distributed under the terms of the Creative Common Attribution 4.0 International License and can be downloaded from the following link: https://www.dropbox.com/sh/p9e00rxjj9r7cbk/AACj3gtVEy5T74rfP58_BtP6a?dl=0).

4. Linguistic Analysis of CLPD-UE-19 Corpus

This section presents the linguistic analysis of the CLPD-UE-19 Corpus. As reported in [34, 35], various edit operations are performed on the source text to create plagiarized text, particularly when it the source text is reused for paraphrased plagiarism. Below we discuss the various edit operations which we observed while carrying out linguistic analysis on a subset of CLPD-UE-19 Corpus (note that, we used 50 source-suspicious document pairs for the linguistic analysis presented in this section) (Figures 5–9).

4.1. Replacing Pronoun with Noun. In these edit operations, a pronoun is replaced by actual name or vice versa in source and suspicious document, for instance:

Original:

جب مغربی طاقتیں کسی ملک کے وسائل پر قبضہ کرنا چاہتی ہیں تو اس ملک کے حکومت کو مجبور کیا جاتا ہے کہ وہ ملک کے خزانوں کو نجکاری کیلئے مارکیٹ میں لائے۔ تو یہ مغربی بینکرز ان کو اپنی شرائط پر کوڑی کے داموں خرید لیتے ہیں۔تو راتوں رات ملک کے سارے خزانوں پر انکا قبضہ ہوجاتا ہیں۔ اسلامی معاشی نظام میں ملکی خزانوں پر ساری عوام کا حق ہوتا ہے لہذا اسکو فروخت نہیں کیا جاسکتا۔لہذا اسلامی معاشی نظام میں نجکاری کے عمل کو محدود کردیا جائے گا

Artificially Plagiarized Copy:

When the Western powers wish to capture a country's resources for the country's government is forced to denationalize the country's treasures dropped at market. If the worth of West Bankers obtain penny on their own terms as a result of the night they're treasures of the country. monotheism financial set-up and also the public's right to national treasuries, therefore it didn't sell. so privatization method within the monotheism financial set-up would be restricted.

Figure 2: An example of plagiarized document created using artificial paraphrasing approach.

Original:

چین توانائی اور دیگر بہت سے منصوبہ جات میں جن میں سینڈک کا منصوبہ، گوادر پورٹ کا منصوبہ پاکستان کو ریلوے انجن کی فراہمی اور دیگر بے شمار ایسے منصوبہ جات ہیں جن میں پاکستان کو چین کی بھرپور مدد حاصل ہے جس سے پاکستان اور چین دوستی کے ایک ایسے رشتے میں بندھے ہوئے ہیں جس کی شاید ہی پوری دنیا میں کوئی مثال موجود ہو اور شاید نہ ہی کبھی ہوگی

Manually Plagiarized Copy:

There are many projects in which China provided their full aid, these projects not only include energy, but also gawadar port, sindic, Pakistan Railway engine supply, and such are other types of huge projects are part of planning. So, in this way, Pakistan and China have good harmony between them. This instance of friendship will and might not seen on anywhere.

Figure 3: An example of plagiarized document created using manual paraphrasing approach.

Original:

ڈاکٹر سر علامہ محمد اقبال (9 نومبر 1877ء تا 21 اپریل 1938ء) بیسویں صدی کے ایک معروف شاعر، مصنف، قانون دان، سیاستدان، مسلم صوفی اور تحریک پاکستان کی اہم ترین شخصیات میں سے ایک تھے۔ اردو اور فارسی میں شاعری کرتے تھے اور یہی ان کی بنیادی وجہ شہرت ہے۔

Non-Plagiarized

Sir Muhammad Iqbal (9 November 1877 –21 April 1938), widely known as Allama Iqbal was a philosopher, poet, mystic and politician in British India who is widely regarded as having inspired the Pakistan Movement. He is considered one of the most important figures in Urdu literature.

Figure 4: An example of nonplagiarized document from our proposed corpus.

Table 2: Corpus statistics.

Size (count of words)	Level name/plagiarized and nonplagiarized/ plagiarized version (total count)			Subject domains								
				CS	*GT*	*Phy*	*Bio*	*EE*	*Zol*	*Psy*	*PS*	*MS*
≤50	(Small)	NP: 450*		100	50	75				25		
		Plagiarized	AT (300)	100	50	99				51		
			AP (300)	100	50	99				51		
			MP (290)	100	50	90				50		
>50 and ≤100	Paragraph (medium)	NP: 225		50	25			20	75		15	40
		Plagiarized	AT (150)	50	25			15			10	50
			AP (150)	50	25			15			10	50
			MP (148)	50	25			15			10	48
≥100 and ≤200	Essay (large)	NP: 135		30	15				33		57	
		Plagiarized	AT (90)	30	15		45					
			AP (90)	30	15		45					
			MP (70)	30	15		25					
			Total	720	360	363	115	65	108	177	102	188

CS: Computer science, *GT*: General Topics, *Phy*: Physics, *Bio*: Biology, *EE*: Electrical Engineering, *Zol*: Zoology, *Psy*: Psychology, *PS*: Pak Studies, *MS*: Management Sciences (200 nonplagiarized documents are from countries domain).

S: گو کہ انہوں نے اس نئے ملک کے قیام کو اپنی آنکھوں سے نہیں دیکھا لیکن انہیں پاکستان کے قومی شاعر کی حیثیت حاصل ہے۔

D: Iqbal has been a national poet of Pakistan Although, he did not see the establishment of the new country with his own eyes

Figure 5: An example of replacing pronoun with noun.

S: ایک چھوٹا سا مکان لے کر اس میں رہنے لگے ، مرتے دم تک یہیں رہے

D:He spent the rest of his life in a small house which he took for rent

Figure 6: An example of changing order of text paraphrasing.

S: بین الاقوامی تجارت معاشیات کی ایک شاخ ہے۔ بنیادی طور پر یہ بین الاقوامی معاشیات کی ایک ذیلی شاخ ہے۔

D: In the branch of economics there exist international trade but basically it is a sub-branch of international economics.

Figure 7: An example of changing source text by adding words.

S: پاکستان فوج کا قیام ۱۹۴۷ میں پاکستان کی آزادی پر عمل میں آیا یہ ایک رضاکار پیشہ ور جنگجو قوت ہے۔

D: It was established on August 14, 1947. Brave, volunteer and sacrificing warriors are main features of them.

Figure 8: An example of paraphrasing text by date completion.

S: جیسا کے جال یا ویب کا مفہوم ہے کہ یہ تمام اطراف پھیلا ہوا ہوتا ہے یعنی باالفاظ دیگر ہر طرف رابطے میں ہوتا ہے اسی طرح رابط کا لفظ بھی اسی مفہوم کی ترجمانی کرتا ہے۔

D: Like the meaning of net spread every where so as web

Figure 9: An example of summarizing source text in plagiarized document.

4.2. Order Change with Add/Delete Words. It is also a common approach used in edit operation. In this approach, later part of the source text is quoted first in the suspicious text and vice versa like.

4.3. Continuing Sentences: Adding Words. Combining two sentences by using an additional word is the most used approach in rewriting text, for example.

4.4. Date Completed. It is another approach where an event in the source text is rewritten in context of the event date and place in suspicious document.

4.5. Summary. In this category, an abstract description of the rewritten text in suspicious document is used in place of long narrations in the source document.

The corpus contains a number of examples of order changes and changing active to passive and direct to indirect and vice versa. Such examples reflect that edit operations change the source text so that it is not a verbatim case. It is not an easy case for plagiarism detection.

5. Translation + Monolingual Analysis of CLPD-UE-19 Corpus

For convenience, this section is further divided into three Sections: starting with experimental setup, next two sections describe detailed and comprehensive analysis of the corpus.

5.1. Experimental Setup. To analyze the quality of artificially and manually paraphrased levels of rewritten cases, we applied translation + monolingual analysis approach on our proposed corpus. Using this approach, we automatically translated source documents (in Urdu) into English using Google Translator. Now, both source and suspicious documents are in the same language, i.e., English. After that, we computed mean similarity scores for source-suspicious document pairs for all four categories (automatic translation copy, artificially paraphrased copy, manually paraphrased copy, and nonplagiarized) using n-gram overlap and longest common subsequence approaches.

To compute similarity scores between source-suspicious document pairs, we applied containment similarity measure [36] (equation (1)). Using the n-gram overlap approach, similarity score between source-suspicious document pair is computed by counting common n-grams between two

documents divided by the number of n-grams in both or any one of the documents. If $S(X, n)$ and $S(Y, n)$ represent word n-grams of length n in source and suspicious document, respectively, then similarity between them using containment similarity measure is computed as follows:

$$\text{Scontainment}(X, Y) = \frac{|S(X, n) \cap S(Y, n)|}{|S(X, n)|}. \quad (1)$$

We used another simple and popular similarity estimation model, longest common subsequence (LCS), to compute the mean similarity scores for four levels of rewrite in CLPD-UE-19 Corpus. Using the LCS approach, for a given pair of source-suspicious text (X and Y), we first computed the LCS between source-suspicious strings and then divided the LCS score with the length of smaller document to get a normalized score between 0 and 1 (equation (2)). Note that LCS method is order-preserving, and LCS score is affected by edit operations performed on source text to generate plagiarized text:

$$\text{LCSnorm}(X, Y) = \frac{|\text{LCS}(X, Y\ )|}{\min(|X|, |Y|)}. \quad (2)$$

5.2. Partial (Domainwise) Analysis. This dimension provides us an opportunity for microlevel and size-oriented domain analysis. Size is one of the dimensions in the rewritten cases. For this purpose, few sample documents from different domains have been randomly selected. Automatic translation copy (ATC) of a source document is compared with artificial and manual paraphrased versions of the same document. Bi, tri, and tetragram split has been applied to identify word to the sentence level similarity between different levels of the rewritten text. An empirical based analysis has been carried out for documents related to all the domains, but only results of only three domains for all size documents are listed here. Almost all results showing that n-gram similarity between both levels of rewrite decreases gradually as values of n increase.

5.2.1. Discussion. It is observed that overall average word n-gram similarity in small-sized manually paraphrased copies of documents is less than large- and medium-sized cases similarity. It also reflects that paraphrasing small-sized text using different edit operations is more paraphrased as compared to other sizes of suspicious documents and hence difficult to detect as well.

In Tables 3–5 and Figure 10, it is noteworthy that 4-gram value or even 3-gram value in most of the cases approaches to zero. It reflects that how well a source document has gradually been altered in both APC and MPC levels of rewrite across the entire corpus. Only a few documents out of such a large corpus have high value of similarity between the source and its MPC level because plagiaries have not used any major paraphrasing techniques for rewriting the source text. But, in such a large corpus of more than 2300 documents, these are only a few such cases.

To have a better view of rewriting levels, we apply APC- and MPC-wise average n-gram approach also, the results of which are presented in Table 6. As per Figure 11, the similarity ratio in most of the APC cases is higher than MPC cases. It also indicates that artificial paraphrasing techniques are still slightly not as precise in paraphrasing source text as compared to the manual effort.

Table 3: Comparison of rewrite levels of *medium* documents from *Pak Study* domain.

	MPC			APC		
	2-gram	3-gram	4-gram	2-gram	3-gram	4-gram
Document 0002.txt	0.153	0.042	0	0.625	0.521	0.457
Document 0005.txt	0.110	0.049	0.025	0.659	0.519	0.388
Document 0006.txt	0.143	0.040	0.008	0.587	0.448	0.347
Document 0009.txt	0.114	0.023	0	0.466	0.322	0.209
Document 0011.txt	0.210	0.066	0	0.387	0.262	0.167

Table 4: Comparison of rewrite levels of *large-sized* documents from the *Biology* domain.

	MPC			APC		
	2-gram	3-gram	4-gram	2-gram	3-gram	4-gram
Document-0041.txt	0.111	0.038	0	0.370	0.231	0.120
Document-0042.txt	0.120	0.042	0	0.280	0.042	0
Document-0087.txt	0.324	0.182	0.063	0.588	0.515	0.469
Document-0094.txt	0.455	0.286	0.150	0.364	0.190	0.050
Document-0095.txt	0.381	0.250	0.105	0.429	0.250	0.053

Table 5: Comparison of rewrite levels of *small sized* documents from the *Physics* domain.

	MPC			APC		
	2-gram	3-gram	4-gram	2-gram	3-gram	4-gram
Document-0066.txt	0.113	0.025	0	0.463	0.329	0.231
Document-0068.txt	0.103	0.026	0	0.449	0.234	0.105
Document-0070.txt	0.218	0.091	0.066	0.487	0.338	0.211
Document-0072.txt	0.121	0.031	0	0.803	0.708	0.609
Document-0075.txt	0.133	0.068	0.014	0.547	0.419	0.329

5.3. Complete (Corpus-Based) Analysis. Table 7 shows the mean similarity scores obtained using n-gram overlap and LCS approaches. AT refers to automatic translation, APC refers to artificial paraphrased copy, MPC refers to manually paraphrased copy, and NP refers to nonplagiarized. 1-gram refers to mean similarity scores generated using n-gram overlap approach, where $n = 1$ (i.e., unigram). Similarly, 2-gram refers to mean similarity scores generated using n-gram overlap approach, where $n = 2$ (i.e., bigram) and so on. Mean similarity scores obtained using LCS approach are referred as LCS. Note that mean similarity score for AT is 1.00 for all methods. The reason is that we used Google Translator for both creating AT cases of plagiarism (Section 3.2) and M + TA analysis (presented in this section). Therefore, the two translations are exactly same generating a similarity score of 1.00 for AT.

As expected, similarity score drops as the level of rewrite increases (from AT to NP). This shows that it is hard to

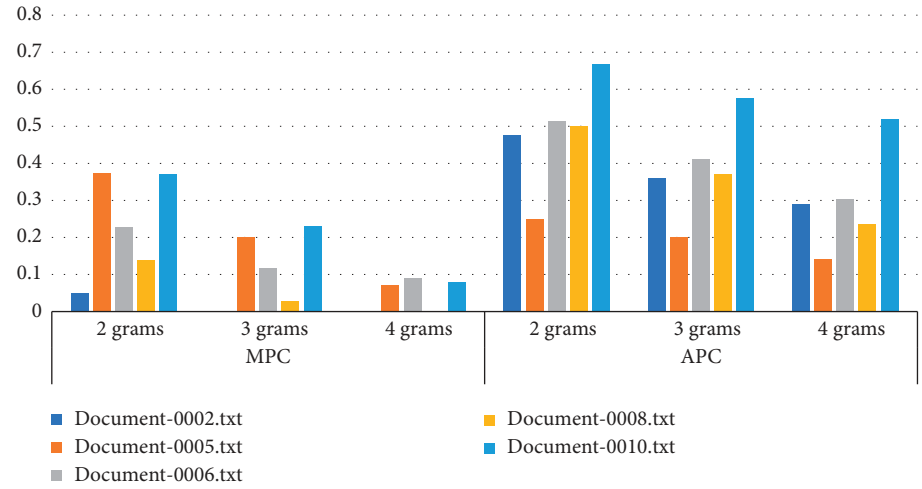


Figure 10: Comparison of manually paraphrased copy (MPC) and artificially paraphrased copy (APC) based on small-sized documents from the Psychology domain.

Table 6: Rewrite level-wise averaged *n*-gram-based small-sized documents from the Psychology domain.

Documents/rewrite levels	MPC	APC
Document-0002.txt	0.017	0.374
Document-0005.txt	0.215	0.198
Document-0006.txt	0.146	0.41
Document-0008.txt	0.056	0.369
Document-0010.txt	0.227	0.588

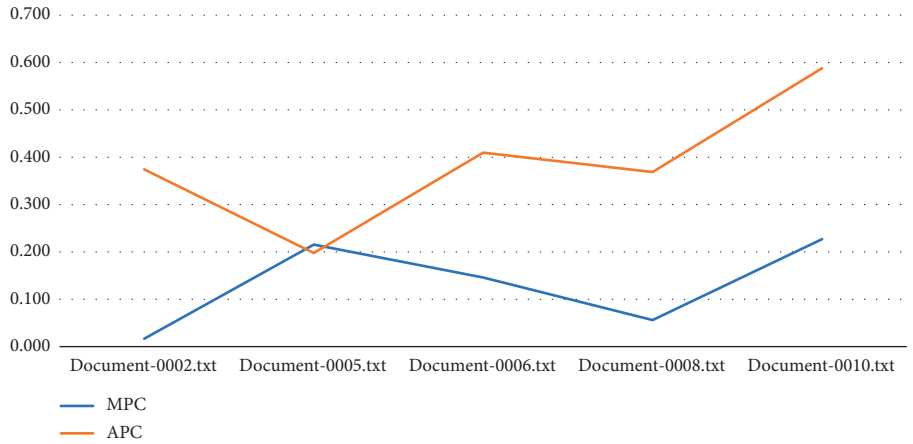


Figure 11: Averaged *n*-gram overlap scores for manually paraphrased copy (mpc) and artificially paraphrased copy (APC) documents.

Table 7: Mean similarity scores for four levels of rewrite in the CLPD-UE-19 Corpus using *n*-gram overlap and LCS approaches.

Method\rewrite levels	At	APC	MPC	NP
1-gram	1.00	0.68	0.52	0.22
2-gram	1.00	0.44	0.21	0.01
3-gram	1.00	0.31	0.11	0.00
4-gram	1.00	0.22	0.07	0
5-gram	1.00	0.16	0.04	0
LCS	1.00	0.20	0.15	0.05

detect plagiarism when the level of rewrite increases. This also shows that suspicious documents in the CLPD-UE-19 Corpus are generated using different obfuscation strategies. For *n*-gram overlap approach, mean similarity scores drops as the length of *n* increases, indicating that it is hard to find long exact matches in the source-suspicious document pairs. For LCS approach, the score is quite low compared to 1-gram approach. This highlights the fact that the order of texts in the source and suspicious document pair is significantly different which makes it hard to find longer matches.

6. Conclusion

The main goal of this study was to develop a large benchmark corpus of cross-lingual cases of plagiarism for Urdu-English language pair at four levels of rewrite including automatic translation, artificial paraphrasing, manual paraphrasing, and nonplagiarized. There are total 2,398 document pairs in our proposed corpus: 1,588 are plagiarized and 810 are nonplagiarized. Plagiarized documents are created using three obfuscation strategies: automatic translation (540 documents), artificial paraphrasing (540 documents), and manual paraphrasing (508 documents). Wikipedia articles are used as source texts and categorized into small, medium, and large documents. Crowdsourcing approach has been applied to create our proposed corpus. We also performed linguistic analysis and translation + monolingual analysis of our proposed corpus. Our empirical analysis showed that there is a clear distinction in four levels of rewrite in our proposed corpus, which makes the corpus more realistic and challenging. Being an emerging area of research [37], in future, we plan to apply cross-lingual plagiarism detection techniques on our proposed corpus.

Data Availability

The authors declare that the data mentioned and discussed in this paper will be provided, if required.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are thankful to all the volunteers for their valuable contribution in construction of the CLPD-UE-19 Corpus.

References

[1] A. Barrón-Cedeno, P. Rosso, E. Agirre, and G. Labaka, "Plagiarism detection across distant language pairs," in *Proceedings of the 23rd International Conference on Computational Linguistics Association for Computational Linguistics*, pp. 37–45, Beijing, China, August 2010.

[2] A. Barrón-Cedeño, P. Gupta, and P. Rosso, "Methods for cross-language plagiarism detection," *Knowledge-Based Systems*, vol. 50, pp. 211–217, 2013.

[3] B. Stein and S. M. zu Eissen, "Intrinsic plagiarism analysis with meta learning," in *Proceedings of the PAN 2007*, p. 276, Amsterdam, Netherlands, July 2007.

[4] B. Martin, "Plagiarism: a misplaced emphasis," *Journal of Information Ethics*, vol. 3, no. 2, p. 36, 1994.

[5] S. Hussain, "Complexity of Asian writing systems: a case study of Nafees Nasta'leeq for Urdu," in *Proceedings of the 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society*, Asian Media Information Center, Singapore, June 2003.

[6] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in *Proceedings of the 23rd international conference on computational linguistics: Association for Computational Linguistics*, pp. 997–1005, Beijing, China, August 2010.

[7] C. H. Lee, C. H. Wu, and H. C. Yang, "A platform framework for cross-lingual text relatedness evaluation and plagiarism detection," in *Proceedings of the 3rd International Conference on Innovative Computing Information and Control ICICIC'08*, p. 303, Dalian, China, June 2008.

[8] Z. F. Alfikri and A. Purwarianti, "The construction of Indonesian-English cross language plagiarism detection system using fingerprinting technique," *Jurnal Ilmu Komputer dan Informasi*, vol. 5, no. 1, pp. 16–23, 2012.

[9] A. Aljohani and M. Mohd, "Arabic-English cross-language plagiarism detection using winnowing algorithm," *Information Technology Journal*, vol. 13, no. 14, pp. 2349–2355, 2014.

[10] H. Asghari, K. Khoshnava, O. Fatemi, and H. Faili, "Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus," in *Proceedings of the CLEF 2015*, Toulouse, France, September 2015.

[11] R. Kothwal and V. Varma, "Cross lingual text reuse detection based on keyphrase extraction and similarity measures," in *Multilingual Information Access in South Asian Languages*, pp. 71–78, Springer, Berlin, Germany, 2013.

[12] M. El-Haj, U. Kruschwitz, and C. Fox, "Creating language resources for under-resourced languages: methodologies, and experiments with Arabic," *Language Resources and Evaluation*, vol. 49, no. 3, pp. 549–580, 2015.

[13] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "A multilingual, multi-style and multi-granularity dataset for cross-language textual similarity detection," in *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia, May 2016.

[14] P. E. Koehn, "A parallel corpus for statistical machine translation," in *Proceedings of the MT summit*, vol. 5, pp. 79–86, Phuket, Thailand, September 2005.

[15] C. K. Kent and N. Salim, "Web based cross language plagiarism detection," in *Proceedings of the Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSIM)*, pp. 199–204, Bali, Indonesia, September 2010.

[16] M. Franco-Salvador, P. Rosso, and M. Montes-y-Gómez, "A systematic study of knowledge graph analysis for cross-language plagiarism detection," *Information Processing & Management*, vol. 52, no. 4, pp. 550–570, 2016.

[17] M. L. Littman, S. T. Dumais, and T. K. Landauer, "Automatic cross-language information retrieval using latent semantic indexing," in *Cross-Language Information Retrieval*, pp. 51–62, Springer, Boston, MA, USA, 1998.

[18] K. Kishida, "Technical issues of cross-language information retrieval: a review," *Information Processing & Management*, vol. 41, no. 3, pp. 433–455, 2005.

[19] M. Hagen, M. Potthast, and B. Stein, "Source retrieval for plagiarism detection from large web corpora: recent approaches," in *Proceedings of the CLEF 2015*, Toulouse, France, September 2015.

[20] E. Stamatatos, M. Potthast, F. Rangel, P. Rosso, and B. Stein, "Overview of the PAN/CLEF 2015 evaluation lab," in *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 518–538, Toulouse, France, September 2015.

[21] 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.

[22] R. Steinberger, B. Pouliquen, A. Widiger et al., "The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages," https://arxiv.org/abs/cs/0609058.

[23] Z. Ceska, M. Toman, and K. Jezek, "Multilingual plagiarism detection," in *Artificial Intelligence: Methodology, Systems, and Applications*, pp. 83–92, Springer, Berlin, Germany, 2008.
[24] I. Hanif, R. M. A. Nawab, A. Arbab, H. Jamshed, S. Riaz, and E. U. Munir, "Cross-language Urdu-english (CLUE) text alignment corpus," in *Proceedings of the CLEF 2015*, Toulouse, France, September 2015.
[25] A. J. Head and M. B. Eisenberg, "How today's college students use wikipedia for course-related research," *First Monday*, vol. 15, no. 3, 2010.
[26] I. Martınez, "Wikipedia usage by Mexican students. The constant usage of copy and paste," in *Proceedings of the Wikimania 2009*, Buenos Aires, Argentina, August 2009.
[27] A. Barrón-Cedeno, P. Rosso, S. L. Devi, P. Clough, and M. Stevenson, "Pana fire: overview of the cross-language Indian text re-use detection competition," in *Multilingual Information Access in South Asian Languages*, pp. 59–70, Springer, Berlin, Germany, 2013.
[28] G. Judge, "Plagiarism: bringing economics and education together (with a little help from it)," *Computers in Higher Education Economics Reviews (Virtual edition)*, vol. 20, pp. 21–26, 2008.
[29] D. L. McCabe, "Cheating among college and university students: a North American perspective," *International Journal for Educational Integrity*, vol. 1, no. 1, 2005.
[30] C. Park, "In other (people's) words: plagiarism by university students–literature and lessons," *Assessment & Evaluation in Higher Education*, vol. 28, no. 5, pp. 471–488, 2003.
[31] P. Clough and M. Stevenson, "Developing a corpus of plagiarised short answers," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 5–24, 2011.
[32] J. Nair, K. A. Krishnan, and R. Deetha, "An efficient English to Hindi machine translation system using hybrid mechanism," in *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2109–2113, Jaipur, India, September 2016.
[33] E. M. Balk, M. Chung, M. L. Chen, T. A. Trikalinos, and L. K. W. Chang, "Assessing the accuracy of google translate to allow data extraction from trials published in non-english languages," Agency for Healthcare Research and Quality (US), Rockville, MD, USA, Report no: 12(13)-EHC145-EF, 2013.
[34] M. Vila, M. A. Martí, and H. Rodríguez, "Is this a paraphrase? What kind? Paraphrase boundaries and typology," *Open Journal of Modern Linguistics*, vol. 4, no. 1, pp. 205–218, 2014.
[35] M. Sharjeel, R. M. A. Nawab, and P. Rayson, "COUNTER: corpus of Urdu news text reuse," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 777–803, 2017.
[36] R. M. A. Nawab, *Mono-lingual paraphrased text reuse and plagiarism detection*, University of Sheffield, Sheffield, England, Ph.D. thesis, 2012.
[37] S. Sameen, M. Sharjeel, R. M. A. Nawab, P. Rayson, and I. Muneer, "Measuring short text reuse for the Urdu language," *IEEE Access*, vol. 6, pp. 7412–7421, 2018.