Develop corpora and methods for cross-lingual text reuse detection for English Urdu language pair at lexical, syntactical, and phrasal levels

Iqra Muneer1,2 · Rao Muhammad Adeel Nawab1

Accepted: 2 August 2022 / Published online: 9 September 2022 © The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

In recent years, Cross-Lingual Text Reuse Detection (CLTRD) has attracted the attention of the research community because large digital repositories and efficient Machine Translation systems are readily and freely available, which makes it easier to reuse text across the languages and very difficult to detect it. In the previous studies, the problem of CLTRD for the English-Urdu language pair has been explored at the sentence/passage and document level, and benchmark corpora and methods have been developed. However, there is a lack of benchmark corpora and methods for the CLTRD for the English-Urdu language pair at the lexical, syntactical, and phrasal levels. To fulfill this research gap, this study presents three large benchmark corpora for detecting the Cross-Lingual Text Reuse (CLTR) at three levels of rewrite (Wholly Derived (WD), Partially Derived (PD), and Non Derived (ND)). The CLEU-Lex, CLEU-Syn and CLEU-Phr corpora contain 66,485 (WD = 22,236, PD = 20,315 and ND = 23,934), 60,267 (WD = 20,007, PD = 16,979 and ND = 23,281) and 60,106 (WD = 23,862, PD = 15,878 and ND = 20,366) CLTR pairs respectively. As a secondary major contribution, we have applied the Cross-Lingual Word Embedding (CLWE), Cross-Lingual Semantic Tagger (CLST), and Cross-Lingual Sentence Transformer (CLSTR) based methods on our three proposed corpora for the CLTRD. Our extensive experimentation showed that for the binary classification task, the best results on the CLEU-Lex corpus were obtained using the cross-lingual sentence transformer ($F_1 = 0.80$). For the CLEU-Syn and CLEU-Phr corpora, the best results were obtained using the cross-lingual sentence transformer and a combination of the CLWE, CLST and CLSTR methods ($F_1 = 0.92$ on CLEU-Syn and $F_1 = 0.94$ on CLEU-Phr). For the ternary classification task, the best results on the CLEU-Lex corpus were obtained using the cross-lingual sentence transformer method ($F_1 = 0.69$). For the CLEU-Syn corpus, the best results were obtained using a combination of the CLWE, CLST, and CLSTR methods ($F_1 = 0.82$). For the CLEU-Phr corpus the best results were obtained using cross-lingual sentence transformer and combination of CLWE, CLST, and CLSTR methods ($F_1 = 0.82$).
To foster and promote research in Urdu (a low-resourced language) all the three proposed corpora are free and publicly available for research purposes.

**Keywords** Cross-lingual text reuse · English-Urdu language pair · Lexical · Syntactical · Phrasal · Cross-lingual word embedding · Cross-lingual semantic tagger · Cross-lingual sentence transformer

### 1 Introduction

Cross-Lingual Text Reuse (CLTR) occurs when the text(s) in one language (L1) is reused to create a new text in another language (L2). In recent years, it has become relatively very easy to reuse text across languages due to the free and easy access of digital repositories containing articles on the same topic in multiple languages (e.g., Wikipedia¹), and the availability of the efficient Machine Translation systems (e.g., Google Translation² and Bing Translator³). CLTRD has many potential applications such as Cross-Lingual Question Answering (CLQA), Cross-Lingual Plagiarism Detection (CLPD), and Cross-Lingual Information Retrieval (CLIR) Ferrero et al. (2017).

The task of Cross-Lingual Text Reuse Detection can be broadly categorized into two categories: (1) Cross-Lingual Local Text Reuse Detection (CLLTRD) (2) Cross-Lingual Global Text Reuse Detection (CLGTRD) (Sameen et al. 2017). In the former case, the new text (in L2) is created by reusing words, sentences, or passages from the source(s) (in L1), whereas in the letter case, the new text (in L2) is created by reusing entire source document(s) (in L1). In both CLLTRD and CLGTRD, the cross-lingual cases of text reuse and be either: (1) cross-lingual artificial cases of text reuse - which are created automatically by using automatic translation and text rewriting tools, (2) cross-lingual simulated cases of text reuse—which are manually created by asking humans to take a text (in L1) and create its reused version (in L2), and (c) cross-lingual real cases of text reuse—which are manually created by journalists to create newspaper stories (in L2) by reusing news agencies text(s) (in L1). The main focus of this study is to create three benchmark corpora for the Cross-Lingual Local Text Reuse Detection containing cross-lingual simulated cases of text reuse.

In previous studies, the majority of the efforts have been made to develop different corpora and methods for the CLTRD and the CLPD for a variety of languages paired with English (e.g. English-German Franco-Salvador et al. (2016), English-Russian Bakhteev et al. (2019), English-Indonesian Alfikri and Purwarianti (2012), English-Spanish Potthast et al. (2011), English-Hindi Kothwal and Varma (2013), English-Arabic Aljohani and Mohd (2014), English-Persian Asghari et al. (2015); Hadgu (2018), multi-lingual (English, German, Catalan, Slovene, Spanish, English-Turkish)

---

¹ [https://www.wikipedia.org](https://www.wikipedia.org): Last visited: 10-02-2021.
² [https://translate.google.com](https://translate.google.com): Last visited: 10-02-2021.
³ [https://www.bing.com/translator](https://www.bing.com/translator): Last visited: 10-02-2021.
Štajner and Mladenić (2019), and English-Arabic, English-Spanish and English-Turkish Li et al. (2018), English-Urdu Muneer et al. (2019); Muhammad (2020)). As can be noted from the literature that the CLTRD benchmark corpora and methods for the English-Urdu language pair have been developed at the sentence/passage level Muneer et al. (2019) and document levels Muhammad (2020). However, there is a lack of benchmark CLTRD corpora for the lexical, syntactical, and phrasal levels for the English-Urdu language pair. Therefore, there is a need to develop gold-standard corpora at lexical, syntactical, and phrasal levels for English-Urdu language pair.

The two main objectives of this study are: (1) to develop large benchmark corpora that contain simulated cases of the cross-lingual text reuse at the lexical, syntactical, and phrasal level, and (2) to develop, evaluate, and compare CLTRD methods on our proposed corpora. For the first objective, we have developed three benchmark corpora for the CLTRD for the English-Urdu language Pair: (1) CLEU-Lex, (2) CLEU-Syn, and (3) CLEU-Phr. In all the three proposed corpora, the source text is in English and reused text is in Urdu. The CLEU-Lex, CLEU-Syn and CLEU-Phr corpora comprise of 66,485 (WD = 22,236, PD = 20,315 and ND = 23,934), 60,267 (WD = 20,007, PD = 16,979 and ND = 23,281) and 60,106 (WD = 23,862, PD = 15,878 and ND = 20,366) CLTR pairs respectively. For the second objective, we have proposed and applied three novel methods including, Cross-Lingual Word Embedding, Cross-Lingual Semantic Tagger, and Cross-Lingual Sentence Transformers on our proposed CLTRD corpora for the English-Urdu language pair. We believe that our proposed corpora will be helpful in: (1) making a direct comparison to the existing CLTRD methods for the English-Urdu language pair, (2) developing, evaluating, and comparing new methods for the English-Urdu language pair, (3) developing Cross-lingual dictionaries for the English-Urdu language pair and (4) fostering research in an under-resourced language i.e., Urdu.

The rest of this paper is organized as follows: Sect. 2 discusses existing corpora and methods for the CLTRD. Section 3 presents the corpus generation process used to create the CLEU proposed corpora including the CLEU-Lex, CLEU-Syn, and CLEU-Phr. Section 5 describes the proposed methods for the CLTRD. Section 6 describes the experimental setup. Section 7 presents results and their analysis. Finally, Sect. 8 concludes the paper with future research directions.

2 Related work

In previous studies, lot of efforts have been made to develop different corpora and methods for measuring the CLTRD and CLPD. One of the prominent efforts in this regard is a series of three PAN International Competitions on the CLPD for the English-Spanish and English-German language pairs Potthast et al. (2011). The outcome of these International Competitions is a set of three large benchmark corpora

---

4 https://pan.webis.de Last visited: 10-2-2021.
The PAN-PC corpora are created by using the cross-lingual artificial and simulated cases of text reuse. In all three PAN-PC corpora, the source text is in English and the reused text is in either German or Spanish.

The JRC-EU Corpus and Fairy Tale Corpus Kent and Salim (2010) are the most common cross-lingual corpora for the CLPD task. In another effort, Ceska et al. created the JRC-EU and the Fairy Tale corpora for the CLPD task Ceska et al. (2008). The JRC-EU corpus contains 400 documents from the legislative reports of the European Union with 200 English source documents and 200 Czech documents Potthast et al. (2011). Fairy-tale corpus contains 54 documents as 27 source documents in English, and 27 suspicious documents in Czech). Ceska et al. applied the MLPlag method based on EuroWordNet thesaurus and obtained ($F_1 = 72.53\%$ and $F_1 = 100\%$) scores on the JRC-EU and Fairy Tale Corpus respectively. Both corpora are not freely available to download.

Barrón-Cedeño et al. proposed a CLTRD corpus to detect the CLTR at the document level for the English-Hindi language pair (called Cross-Language Indian Text Reuse (CLITR) corpus) Barrón-Cedeno et al. (2013). The proposed corpus contains 5,032 source documents in English, and 388 reused documents in Hindi respectively. In this corpus, all text pairs are based on artificial and simulated cases of the CLTR, and have been annotated at four levels (exact copy = 79, light revision = 99, heavy revision = 98, and original = 112). To develop and evaluate the CLTRD systems for the English-Hindi language pair, the CLITR corpus was presented in an International Competition. The best system in the competition was based on keyphrase extraction Kothwal and Varma (2013) obtained ($F_1 = 0.79$). This corpus is freely available for download.

Haneef et al. proposed a benchmark corpus to measure the CLPD at the document level (called Cross-Language Urdu English Text Alignment Corpus (CLUE-TAC)) Hanif et al. (2015). The corpus consists of 500 source documents in English, and 500 reused in Urdu, and document pairs respectively based on simulated cases. The corpus consists of three levels of obfuscation (“Near Copy,” “Light Revision,” and “Heavy Revision”), whereas 230 documents in the corpus are “Non plagiarized.” The corpus was used for the development and evaluation of the CLPD systems for the English-Urdu language pair for the text alignment task only as described by PAN organizers Haneef et al. (2019).

Recently, Haneef et al. proposed a large benchmark corpus to measure the CLPD at the document level for the English-Urdu language pair Haneef et al. (2019). The corpus contains 2,395 source in English - suspicious in Urdu document pairs, based on simulated cases of the CLPD. The corpus consists of (automatic translation = 540, artificially paraphrased = 539, manually paraphrased = 508, and Non plagiarized = 808). The author applied the n-gram Overlap and the longest common

---

5 PAN-PC corpora are freely available to download https://www.uni-weimar.de/en/media/structure/. Last visited: 10-2-2021
6 FIRE 2013 competition https://dl.acm.org/doi/proceedings/10.1145/2701336 Last visited: 10-2-2021.
7 https://www.uni-weimar.de/medien/webis/events/panfire-11/panfire11-web/#corpus Last visited: 10-2-2021.
sub-sequence methods, and the best results were obtained using the n-gram Overlap (unigrams) mean similarity scores as (1.00, 0.68, 0.52, and 0.22) for automatic translation, artificially paraphrased, manually paraphrased, and Non plagiarized documents respectively.

Muneer et al. proposed a CLTRD corpus for the CLTRD at the sentence/passage level for English-Urdu language pair (called CLEU corpus) Muneer et al. (2019). The proposed corpus comprises of 3,235 real source in English - reused in Urdu, CLTR pairs of journalism, which are manually annotated into three classes as (Near Copy = 751, Paraphrased Copy = 1751, and Independently Written = 733). To develop and evaluate the CLTRD systems for the English-Urdu language pair, three sets of methods (N-gram Overlap, Greedy String Tiling, Longest Common Sub-sequence) using T+MA were applied on their proposed CLEU sentence/passage corpus. The best results were obtained (\(F_1 = 0.732\)) using the N-gram Overlap(unigram) and (\(F_1 = 0.552\)) using Greedy String Tiling (GST-mm1) for the binary and ternary classification tasks respectively.

Recently, Sharjeel et al. has proposed a CLTRD corpus at the document level for the English-Urdu language pair (called TREU corpus) Muhammad (2020). The proposed corpus comprises of 2257 CLTR document pairs, source in English - reused in Urdu. The corpus is manually annotated into three classes as (Wholly Derived = 672, Partially Derived = 888, and Non Derived = 697) based on the real cases. The author applied the N-gram Overlap, Greedy String Tiling, Longest Common Sub-sequence, mono-lingual word embedding methods, and the mono-lingual sentence embedding methods based on T+MA. The best results were obtained (\(F_1 = 0.78\)) using the N-gram Overlap (unigram) and (\(F_1 = 0.66\)) using all combine methods for the binary and ternary classification tasks respectively.

Bakhteev et al. proposed the CLPD system to measure the CLPD for the English-Russian language at the document level Bakhteev et al. (2019). The training dataset used for the proposed system consists of 30 million parallel pairs based on the artificial cases of the CLP from Russian and English Wikipedia articles. The best results were obtained (\(F_1 = 0.87\)) using the Translation plus mono-lingual Analysis approach.

Recently, Muneer et. al. Muneer and Nawab (2021) have proposed new methods for the CLTRD for the English-Urdu language pair at the sentence/passage level. The authors have proposed, and compared T+MA based methods using the probabilistic, word embedding, semantic, and deep learning methods. The best performance was reported using ‘Comb-All’ method with \(F_1 = 0.77\), and \(F_1 = 0.61\) for the binary, and ternary classification tasks respectively.

Gupta et al. Gupta et al. (2012) explored three different methods for the problem of duplicate document identification for the two language sets. They explored the Cross-Language Alignment-based Similarity Analysis (CL-ASA), Cross-Language Character N-Grams (CL-CNG), and proposed a concept-based similarity model. The problem was explored for the English-German and the English-Spanish utilizing the Eurovoc conceptual thesaurus which contains 6,797 multilingual concepts for 22 languages. The results showed that their proposed model outperformed the other methods with an average distribution of 0.95.
Franco et al. Franco-Salvador et al. (2016) explored different methods including the Cross-Language Knowledge Graph Analysis (CL-KGA) and the external-data composition neural network (XCNN) for the CLPD for the Spanish-English (ES-EN) and the German-English (DE-EN) language pairs. The best results were obtained using XCNN with plagdet = 0.644, precision = 0.556, granularity = 1.00 and recall = 0.95.

Vstajner et al. Štajner and Mladenić (2019) explored the task of cross-lingual similarity estimation for the English, German, Catalan, Slovene, Spanish, and Croatian versions of Wikipedia. The authors compared cross-lingual latent semantic indexing, low-rank canonical correlation analysis, and a nonlinear bilingual translation model using the monolingual word embedding and kernel approximation. The best results precision = 0.89 were obtained using a word embedding and kernel approximation for cross-lingual similarity detection.

Chang et al. Chang et al. (2020) proposed a new approach for the CLPD called cross-lingual word mover distance for the English-Chinese language. The authors compared the new approach with other multiple methods including the Average, TF-IDF weight, smooth inverse frequency, and T+MA. These methods were compared on two corpora, namely the NDLTD-Paragraph (12,704 pairs), and NDLTD-Sentence (1,772 pairs) at the Paragraph level and sentence level respectively. The best results were obtained using the cross-lingual word mover distance (Hit@1 = 0.9073, Hit@5 = 0.9615, and Hit@10 = 0.9709) for the NDLTD-Paragraph. However, the best results for the NDLTD-Sentence were obtained (Hit@1 = 0.5968) using T+MA, and (Hit@5 = 0.8096, and Hit@10 = 0.8609) using the cross-lingual word mover distance.

Alberto et. al. Barrón-Cedeno et al. (2010) compared three different cross-lingual methods including the CL-ASA, CL-CNG, and T+MA for the CLPD for the English-Basque and Spanish-Basque language pair. The best results were obtained using the T+MA method (recall = 0.77) for the English-Basque and (recall = 0.89) for the Spanish-Basque. They demonstrated that the CL-CNG is more suitable to use when two languages have similar scripts.

Alberto et. al. Barrón-Cedeño et al. (2013) applied three different cross-lingual methods including the CL-ASA, CL-CNG, and T+MA for the CLPD for the Spanish–English language pair. The best results were obtained using T+MA approach recall = 0.51, precision = 0.67, and $F_1 = 0.57$.

Potthast et al. Potthast et al. (2011) explored three different similarity estimation methods for CLPD. These models include the CL-ASA, CL-CNG, and the Cross-Language Explicit Semantic Analysis (CL-ESA). The models were evaluated on 120,000 test documents that were selected from the JRC-Acquis and the Wikipedia corpora for the English, Spanish, French, German, Dutch, and Polish languages. The best results were obtained using the Cross-Language Character N-Grams (CL-CNG) approach with recall = 0.99.

Flores et. al. Flores Sáez et al. (2015) applied different cross-lingual methods including the Cross-Lingual Latent Semantic Analysis (CL-LSA), CL-CNG, CL-ESA, CL-ASA, Pseudo-Cognateness (CL-COG), and the Word Count Ratio for the task of cross-language source code reUse detection for C–Java, Java–Python, and C–Python language pairs. The best results were obtained using the CL-LSA method.
Develop corpora and methods for cross-lingual text reuse...

$F_1 = 0.753$ for the C–Java, $F_1 = 0.861$ for the Java–Python, and $F_1 = 0.444$ for the C–Python language pair.

To summarize, the problem of CLTRD and CLPD have been explored for a variety of languages paired with English language including (e.g. English-German Franco-Salvador et al. (2016) English-Russian Bakhteev et al. (2019), English-Spanish Potthast et al. (2011); Li et al. (2018), English-Hindi Kothwal and Varma (2013), English-Czech, Ceska et al. (2008), and English-Urdu Muneer et al. (2019); Muhammad (2020). The existing corpora contain the artificial, simulated, and real cross-lingual cases of the text reuse and the plagiarism at the sentence, passage, or document levels. For the English-Urdu language pair, the problem of CLTRD has been explored at the sentence/passage and document levels. However, the problem of CLTR has not been explored at the lexical, syntactical, and phrasal levels for the English-Urdu language pair. To fulfill this gap, this study presents three large benchmark corpora containing cross-lingual simulated cases of the text reuse at the lexical, syntactical, and phrasal levels for the English-Urdu language pairs. In addition, we applied the CLWE, CLST, and CLSTR methods on our proposed corpora. As far as we are aware, the three large benchmark corpora for the CLTRD for the English-Urdu language pair proposed in this study have not been previously reported.

3 Corpora generation process

This section presents the process used to create our three proposed cross-lingual corpora including data collection, annotation process (annotation guidelines, annotations, and the Inter-Annotator Agreement), corpora characteristics, and examples from the proposed cross-lingual corpora. Below we describe the proposed corpora generation process in detail.

3.1 Data collection

To create our proposed Cross-Lingual Text Reuse Detection (CLTRD) corpora at the lexical, syntactical and phrasal level, we have used a subset of data from the PPDB 2.0 corpus.\(^8\)Ganitkevitch et al. (2013). In previous studies, these corpora have been used for various tasks including the automatic paraphrase generation and detection Ganitkevitch et al. (2013), sentential paraphrasing as Black-Box Machine Translation Napoles et al. (2016), paraphrase database simplification Pavlick and Callison-Burch (2016), and the compositional paraphrase modeling Wieting et al. (2015).

The PPDB 2.0 corpus comprises of multiple corpora (in multiple languages) at the lexical (PPDB-2.0-s-lexical corpus), syntactical (PPDB-2.0-s-syntactical), and phrasal (PPDB-2.0-s-phrasal). All text pairs in the PPDB 2.0 pair corpus are a paraphrase of each other. There are 2,31,680, 27,89,218, and 10,04,112 text pairs in the

---

\(^8\) the ParaPhrase DataBase http://paraphrase.org/. Last visited: 26-12-2020
English language in the PPDB-2.0-s-lexical corpus, PPDB-2.0-s-syntactical corpus, and PPDB-2.0-s-phrasal corpus respectively.

To develop our proposed Cross-Lingual English Urdu Lexical (CLEU-Lex), Cross-Lingual English Urdu Syntactical (CLEU-Syn), and the Cross-Lingual English Urdu Phrasal (CLEU-Phr) corpora, we have extracted 10,000 text pairs from the PPDB-2.0-s-lexical corpus, 10,000 from the PPDB-2.0-s-syntactical corpus, and 10,000 from the PPDB-2.0-s-phrasal corpus.

The subsets of text pairs extracted from the PPDB-2.0 corpora are in the English language, i.e., mono-lingual. To create cross-lingual corpora at the lexical, syntactical, and phrasal level, a linguist expert (who had expertise in both English and Urdu languages) manually translated both the English texts in each text pair into Urdu. After manual translation, the linguistic expert created different possible paraphrases of Urdu texts. After that, we paired the English and Urdu texts to make English-Urdu text pairs for creating our Cross-Lingual English Urdu corpora at the lexical, syntactical, and phrasal level. After pairing English-Urdu texts, we have obtained 66,485, 60,267, and 60,106 cross-lingual text pairs at the lexical, syntactical, and phrasal level.

3.2 Annotation process

The annotation process is divided into three main steps: preparation of annotation guidelines, (2) annotations, and (3) the calculation of the Inter-Annotator Agreement.

3.2.1 Annotation guidelines

Our main goal is to develop cross-lingual text reuse detection corpora with three levels of rewrite: (1) Wholly Derived - when both texts are near or exact copy of each other, (2) Partially Derived - when both texts are paraphrase of one another, and (3) Non Derived - When both texts are entirely different or unrelated. To achieve this goal, each cross-lingual text pair was manually classified into one of the three categories: (a) Wholly Derived (WD), (b) Partially Derived (PD), or (c) Non Derived (ND), depending upon the relationship between them. To classify a cross-lingual text pair into one of the three categories, the following guidelines were prepared:

(1) Wholly Derived: If Text 02 (Urdu) is an exact translation of Text 01 (English), then that cross-lingual text pair will be annotated as Wholly Derived (WD).
(2) Partially Derived: If Text 02 (Urdu) is a paraphrase of Text 01 (English) then that cross-lingual text pair will be annotated as Partially Derived (PD).
(3) Non Derived: If the content of Text 01 (English) and Text 02 (Urdu) is unrelated then that cross-lingual text pair will be annotated as Non Derived (ND).

The extracted text pairs in the English language can be downloaded from the following link: https://drive.google.com/drive/folders/1RXF6kXytdkHf0Zs-yGpVuJtcXjRV18hF?usp=sharing.
3.2.2 Annotations

Annotation guidelines prepared in the previous step were used to manually annotate the three proposed corpora (CLEU-Lex, CLEU-Syn, and CLEU-Phr). Annotations were carried out by seven annotators A, B, C, D, E, F, G. Annotator A is a post-graduate Natural Language Processing student and Ph.D. scholar in the field of CLTR. All other annotators were under-graduate (7th semester) Natural Language Processing students, native Urdu speakers, with a high level of proficiency in the English. Furthermore, they were also provided with training in the CLTR process and the CLTR edit and rewriting operations with the help of tutorials and state-of-the-art corpora by a domain expert. The main objective of the training was to show the relatedness to different levels of CLTR and cross-lingual text annotation.

Cross-lingual (English Urdu) text pairs for the CLEU-Lex corpus were manually annotated by annotators A, B, and C. For the CLEU-Syn corpus, annotations were carried out by the annotators A, D, and E. Finally, for the CLEU-Phr corpus annotators A, F, and G manually performed the annotations\(^{10}\).

All three proposed corpora were annotated in these steps: In the first step, the first two annotators annotated a subset of 1,000 cross-lingual text pairs. The agreed and conflicting cross-lingual text pairs in the first 1,000 cross-lingual text pairs, were discussed by the annotators and annotation guidelines were revised (if needed). The revised annotation guidelines were used to annotate the full corpus and the inter-annotator agreement was computed for each of each entire corpus. The conflicting pairs of the CLEU-Lex, CLEU-Syn, and the CLEU-Phr were annotated by the Annotator C, E, and G respectively. The Inter-Annotator Agreement was computed for all the proposed corpora.

3.2.3 Inter-annotator agreement

Table 1 shows the detailed statistics of Inter-Annotation Agreements (IAA) and Kappa coefficient for all three cross-lingual corpora. The Inter-Annotator Agreement for the CLEU-Lex corpus, CLEU-Syn corpus, and the CLEU-Phr corpus was 84.0%, 84.6%, and 84.6% respectively. The Kappa coefficient was 76.08%, 77.03%, and 0.78% for the CLEU-Lex corpus, CLEU-Syn corpus, and the CLEU-Phr respectively. As it can be noted that the Inter-Annotator Agreements and Kappa coefficient for all three proposed cross-lingual corpora are good. This highlights the fact that annotation guidelines were very well defined which helped annotators to recognize between different levels of the CLTR in the proposed corpora. In addition, this also shows that the annotators were fully trained and have expertise in the field of CLTR.

In the CLEU-Lex corpus, there are a total of 66,486 instances (52,882 were agreed and 10,504 disagreed between the first two annotators). As can be noted from Table 1, the majority of conflicts are between Partially Derived (PD) and Non Derived (ND) classes, which highlights the fact that the annotators had difficulty in distinguishing between these two classes. The possible reason for this is that a single

---

\(^{10}\) Annotator A is the first author of this paper.
English word may have multiple translations in the Urdu language. The situation gets more complex when an Urdu text is a paraphrase of an English text. Consequently, this makes it more challenging for the annotators to accurately discriminate the PD class from the ND. However, there were only 995 conflicts between the WD and the PD classes, which shows that it was relatively easy to differentiate between these two classes.

In the CLEU-Syn corpus, there are a total of 60,267 instances (51,035 were agreed and 9232 disagreed between the first two annotators). Again, the majority of conflicts are between the Partially Derived (PD) and the Non Derived (ND) classes, which can be seen from Table 1. Again, the reason is the same as a single English word may have multiple translations in the Urdu language. The task will be more complex when the Urdu text is a paraphrase of an English text. Consequently, it becomes more difficult for the annotators to differentiate between these classes. Again, the proportion of conflicted pairs among the WD and the PD is not very high (390 text pairs).

In the CLEU-Phr corpus, there are a total of 60,106 instances (51,474 were agreed and 8632 disagreed between the first two annotators). As can be noted from Table 1, the majority of the conflicts are between the Wholly Derived (WD) and the Partially Derived (PD) classes. The possible reason for this is that it might be difficult to discriminate whether the Urdu and English text pairs are translation or paraphrase of one another. Similar to the CLUE-Syn and CLUE-Lex corpora, the number of conflicts between the PD and ND classes is high (3233 conflicted text pairs).

### 3.3 Corpora characteristics

As can be noted from Table 1, out of 66,485 cross-lingual text pairs, the gold-standard CLEU-Lex contains 22,236 (33.44%) WD, 20,315 (30.5%) PD, and 23,934 (35.99 %) ND cross-lingual text pairs. Similarly, Table 1 shows that out of 60,267 cross-lingual text pairs, 20,007 (0.33%) WD, 16,979 (0.28%) PD, and 23,281 (0.38%) are ND for gold-standard CLEU-Syn. The CLEU-Phr comprises of 60,106 cross-lingual text pairs, out of which 23,862 (0.39%) belong to WD, whereas 15,878 (0.27 %) are from the PD, 20,366 (0.33%) belong to ND class. This indicate that the CLEU-Lex is very well balanced however, the other two corpora are moderately balanced.

Table 2 show detailed statistics of all the proposed corpora. The CLEU-Lex corpus contains in total 66,485 English, and 1,09,016 Urdu tokens, CLEU-Syn consists of 1,42,362 English, and 1,64,375 Urdu tokens, and there are 1,92,126 English, 2,17,952, and Urdu tokens in the CLEU-Phr. This shows that the length of Urdu text is larger than English text in all the proposed corpora. More detailed statistics can be found in Table 2.
All the proposed corpora are standardized in CSV format and these corpora will be freely available to download for research proposes under a Creative Commons CC-BY-NC-SA license\textsuperscript{11}. These CLEU corpora can be accessed from the following link for the reviewers\textsuperscript{12}.

### 3.4 Examples from proposed corpora

Tables 3, 4, and 5, show the WD, PD, and ND cross-lingual text reuse examples from the CLEU-Lex, CLEU-Syn and the CLEU-Phr corpora respectively. As can be noted from Table 3, in the CLEU-Lex corpus, the WD cross-lingual text pairs are an exact translation of another. The PD cross-lingual text pairs are a paraphrase of one another, and finally, the ND cross-lingual text pairs are entirely unrelated to one another. Similarly, Table 4 shows that in the CLEU-Syn corpus, WD cross-lingual text pairs are word to word translations of one another. The PD cross-lingual text pairs show that the English and Urdu texts are paraphrase of one another and ND cross-lingual text pairs are not related to one another. The WD, PD, and ND examples of the CLEU-Phr corpus also show a similar pattern to that of the CLEU-Syn and the CLEU-Lex corpora (see Table 5).

To conclude, in all three proposed CLEU corpora, in the WD class English-Urdu text pairs are almost exact translations of each other, in the PD class Urdu text is a paraphrase of an English text, and in the ND class English-Urdu text pairs are unrelated to one another.

### 4 Baseline method

#### 4.1 Bi-lingual dictionary based method

In the Cross-Language Information Retrieval (CLIR) problem, a bi-lingual dictionary has been used to translate terms or phrases of search queries from one language...
The main purpose of a machine-readable bi-lingual dictionary is to help people to translate texts from one language into another and/or understand foreign-language texts. A machine-readable bi-lingual dictionary is useful in developing systems that contain texts across languages. Moreover, a machine-readable bi-lingual dictionary is beneficial in a variety of natural language processing tasks including cross-lingual machine translation, the CLPD, cross-lingual paraphrase detection, the CLTRD, cross-lingual questioning answering, cross-lingual semantic text similarity, AND cross-lingual duplicate content detection. Using a machine-readable bi-lingual dictionary we can quickly and easily access high-quality translations of words.

| Table 2 | Proposed corpora characteristics |
|---------|----------------------------------|
| Characteristics | CLEU-Lex | CLEU-Syn | CLEU-Phr |
| Total pairs | 66,485 | 60,267 | 60,106 |
| Wholly derived | 22,236 | 20,007 | 23,862 |
| Partially derived | 20,315 | 16,979 | 15,878 |
| Non derived | 23,934 | 23,281 | 20,366 |
| Source | Reused | Source | Reused | Source | Reused |
| Total tokens | 66,485 | 1,09,016 | 1,39,969 | 1,60,624 | 1,84,674 | 1,90,143 |
| Total token (without stop-words) | 66,212 | 77,328 | 67,204 | 56,872 | 88,037 | 57,908 |
| Total types | 7206 | 4842 | 3711 | 3168 | 5223 | 5862 |
| Total types (without stop-words) | 7204 | 5955 | 2538 | 3812 | 3462 | 9302 |
| Min tokens per example | 1 | 1 | 1 | 1 | 1 |
| Max tokens per example | 1 | 4 | 4 | 9 | 7 | 12 |
| Mean of tokens per example | 1 | 2 | 2 | 3 | 3 |
| Median of tokens per example | 1 | 1 | 2 | 2 | 3 |

| Table 3 | Examples of cross-lingual text reuse text pairs from CLEU-Lex corpus |
|---------|----------------------------------|
| Source | Derived | Annotation |
| Situations | حالات | Wholly derived |
| Testifies | گوابی دتا م | Wholly derived |
| Selected | منتخب شده | Wholly derived |
| Risks | خطرات | Wholly derived |
| Situations | مسال | Partially derived |
| Testifies | دیکها | Partially derived |
| Selected | تجویز کرد | Partially derived |
| Risks | خطر ناک | Partially derived |
| Situations | گابیک | Non derived |
| Testifies | چالی | Non derived |
| Selected | ملتو | Non derived |
| Risks | حالات | Non derived |

to another Grefenstette (1998).
For the cross-lingual text reuse detection, the bi-lingual dictionaries can be used to translate terms or phrases from the derived text to the language of source texts. While doing so, context analysis is also considered as we will not only observe the term and its translation are same but are in the same context Daille and Morin (2008).

For our experiments, we used an existing bi-lingual dictionary FA14-MSCS (2016). Each entry of the bi-lingual dictionary contains an Urdu word and its English translation. Given a cross-lingual source-reused text pair from the

| Table 4: Examples of cross-lingual text reuse text pairs from CLEU-Syn corpus |
|---------------------------------|-----------------|-----------------|
| Source                          | Derived         | Annotation      |
| Responded                       | جواب دیا         | Wholly derived |
| Other matters                   | دوسرے معاملات    | Wholly derived |
| Recent years                    | حالیہ برسون      | Wholly derived |
| Unemployment                    | بے روزگاری       | Wholly derived |
| Responded.                      | جواب            | Partially derived |
| Other matters                   | دیگر مسائل      | Partially derived |
| Recent years                    | پچھلے سال        | Partially derived |
| Unemployment                    | روزگاری کے لئے    | Partially derived |
| Responded                       | کے حصول کے لئے    | Non derived |
| Other matters                   | اس کے ایک حصے کے طور پر    | Non derived |
| Recent years                    | قانون سازی کا بنا    | Non derived |
| Unemployment                    | متعلقہ پر   | Non derived |

| Table 5: Examples of cross-lingual text reuse text pairs from the CLEU-Phr corpus |
|---------------------------------|-----------------|-----------------|
| Source                          | Derived         | Annotation      |
| It is clear that                | م: واضح پہ کہ   | Wholly derived |
| I can tell.                     | میں بتا سکتا ہوں | Wholly derived |
| What the hell is going on      | کا مصیبت چل رہی پہ   | Wholly derived |
| Community agency                | برادری کی ابتنس   | Wholly derived |
| What the hell is going on      | م: کا پہ   | Partially derived |
| Community agency                | کمیونٹی پر مبنی تنظیمن | Partially derived |
| Continued support to            | کی حمایت جاری پہ   | Partially derived |
| It is clear that                | م: واضح طور پر پہ   | Partially derived |
| I can tell                      | نتائج   | Non derived |
| What the hell is going on      | اس کے ایک حصے کے طور پر    | Non derived |
| Continued support to            | تھورا اندہ    | Non derived |
| Community agency                | کے ساتھ مکمل طور پر تعمیل | Non derived |
CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora, the baseline bi-lingual dictionary approach was applied as follows. In the first step, both the source and reused texts were tokenized. In the second step, each word in the reused text (Urdu), was translated to English using a bi-lingual dictionary (such that both the source and reused tokens are in the English language). In the third step, the similarity between the source and reused texts was computed using two different similarity coefficients: Overlap similarity coefficient, and Jaccard similarity coefficient.

\[
S_{\text{overlap}} = \frac{|S(S, n) \cap S(T, n)|}{\min(|S(S, n)|, |S(T, n)|)}
\]  
\[
S_{\text{Jaccard}} = \frac{|S(S, n) \cap S(T, n)|}{(|S(S, n)|, |S(T, n)|)}
\]

where S and T represent source and reused (translation) texts respectively.

5 Proposed methods for Cross-Lingual English Urdu Corpora

To demonstrate how the proposed corpora can be used for the development, evaluation, and comparison of the CLTRD systems for the English-Urdu language pair, we have applied different CLTRD methods on our proposed corpora, including the cross-lingual word embedding based method (word embedding based methods), cross-lingual semantic tagger based methods, and the cross-lingual sentence transformer based method. As per our knowledge, these methods have not been previously used for the CLTRD for English-Urdu language pair at the lexical, syntactical, and the phrasal levels. The next sections describe these CLTRD methods in detail.

5.1 Word embedding based methods

5.1.1 Cross-Lingual word embedding based method

The basic idea of word embedding is the representation of the words on their context and words around it Ferrero et al. (2017). The words are represented in a continuous space and those with the same context should be closed in this multi-dimensional space. The word embedding models can be used to measure semantic textual similarity by using the distributed representation of words. Common but efficient (and effective) word embedding architectures include the wor2vec CBOW and skip-gram models Mikolov et al. (2013), Glove Pennington et al. (2014), Ghannay et al. (2016), and canonical correlation analysis (CCA) Upadhyay et al. (2016). These models will
map the words to vectors of real numbers, and follow the instinct that when words are represented, the same words will be represented by the same vector in a common vector space. Word embedding was initially proposed for the mono-lingual comparability analysis, and lately to the cross-lingual word similarity analysis by using a common representation space for more than one language Ruder (2017). Word embedding methods have been used in a range of applications including word sense disambiguation Pelevina et al. (2017), recommendation service Ozsoy (2016), short text similarity Kenter and De Rijke (2015), CLPD Ferrero et al. (2017); Khorsi et al. (2018), and analyzing survey responses or verbatim comments Tang et al. (2014).

We need two pre-trained word embedding models to compute the similarity between the cross-lingual text pairs of the proposed corpora. For this study, we have used the Google pre-trained word embedding model13 Ghannay et al. (2016) for the source (English) text. Similarly, for the reused (Urdu) text, a pre-trained Urdu word embedding model14 Kanwal et al. (2019) has been used.

To compute the similarity between the source-reused text pairs, the cross-lingual word embedding method was used as follows. In the first step, both the source and reused texts were tokenized and pre-processed by removing punctuation marks. In the second step, the pre-trained Google Word2Vec model Ghannay et al. (2016) was used to extract 300 nearest neighbors for all the words in the source text and the Urdu pre-trained Urdu word embedding model Kanwal et al. (2019) was used to extract 300 nearest neighbors for all the words in the reused text. In the next step, the similarity between the word embedding vectors of source and reused texts were computed in two ways: (1) the sum of the word embedding vectors method and (2) the average of the word embedding vectors method.

For the sum of the word embedding vectors method, word embedding vectors of all the source words were summed to get a single source word embedding vector. Similarly, vectors of all the reused words were summed to get a single reused word embedding vector. After that, the similarity between the (summed) source and reused word embedding vectors were computed using the Cosine distance similarity measure (Equation (3)) and the Euclidean distance similarity measure (Equation (4)).

For the average word embedding vectors method, the word embedding vectors of all the source words were averaged to get a single source word embedding vector. Similarly, the vectors of all the reused words were averaged to get a single reused word embedding vector. After that, the similarity between the (averaged) source and reused word embedding vectors were computed using the Cosine similarity measure (Equation (3)) Lahitani et al. (2016) and the Euclidean distance measure (Equation (4)) Vijaymeena and Kavitha (2016).

---

13 Pre-trained Google word embedding model is trained for the English language on 100 billion words from a Google News dataset
14 Urdu word embedding model is trained on MK-PUCIT corpus with 28,006,880 tokens
Where $|S|$ and $|R|$ represents the length of source and reused texts respectively. The Cosine similarity measure allows partial matching, which enables a better estimation of similarity. The Euclidean distance compares the shortest distance between the source and reused texts.

\[ Sim(S, R) = \frac{\vec{S} \cdot \vec{R}}{|S| \times |R|} \]  

(3)

Where $(\vec{S})$ and $(\vec{R})$ represents the source and reused text respectively. The Euclidean distance compares the shortest distance between the source and reused texts.

\[ Euclidean(S, R) = \sqrt{(\vec{S} - \vec{R}) \cdot (\vec{S} - \vec{R})} \]  

(4)

5.2 Semantic similarity based methods

5.2.1 Cross-Lingual semantic tagger based method

To carry out the semantic analysis of the text, Rayson et al. Rayson et al. (2004) have developed the USAS (UCREL Semantic Analysis System) frameworks.15 The USAS framework contains 21 major discourse tags/fields, which are further subdivided into 232 fine-grained tags/sub-fields.16 Using the USAS framework, semantic taggers have been developed for a range of languages including English, Dutch, Chinese, Portuguese, Italian, and Spanish, etc. The USAS semantic annotation tool has been used in a variety of applications including corpus linguistics, electronic dictionaries, software engineering, content analysis Rayson et al. (2004), for the development of the multi-lingual semantic annotation system for the Italian, Chinese and Brazilian languages Piao et al. (2015), and for the automatic extraction of multiword expressions (MWEs) Piao et al. (2005) etc.

To develop the cross-lingual semantic tagger based methods, we need two semantic taggers to compute the similarity between the cross-lingual text pairs of the proposed corpora. For this study, we have used the USAS English semantic tagger for the English text17. However, for the Urdu text, Urdu semantic tagger by Shafi (2019) USAS Urdu tagger contains the single word and multi-word English semantic lexicons, which have 56,318 and 16,871 entries respectively with 21 major semantic fields and 232 sub-fields with USAS (UCREL Semantic Analysis System) is used18. Note that, semantic tagset for English semantic tagger and Urdu semantic tagger are the same.

15 http://ucrel.lancs.ac.uk/usas/, Last visited: 20-12-2020.
16 The full tagset is available at the following link: http://ucrel.lancaster.ac.uk/usas/semtags.txt Last visited: 13-08-2020
17 English semantic tagger can be used online from the following link: http://ucrel-api.lancaster.ac.uk/usas/tagger.html Last visited: 10-02-2021
18 https://raw.githubusercontent.com/UCREL/Multilingual-USAS/master/Urdu/Urdu_Semantic_Lexicon.txt Last visited: 10-02-2021.
For this study, we computed the semantic similarity between the source-reused text pairs as follows. In the first step, USAS online English tagger was used to assign semantic tags to all source text pairs in the Cross-Lingual English Urdu (CLEU) corpora. Similarly, the USAS Urdu tagger was used to assign semantic tags to all Urdu texts in all CLEU corpora. After that, the similarity between the semantic tags of the source-reused text pairs was computed using two different similarity coefficients: Overlap similarity co-efficient, and the Jaccard similarity similarity coefficient.

If $S_{tags}$ and $R_{tags}$ represent the USAS semantic tags assigned to the source and reused texts, then similarity between $S_{tags}$ and $R_{tags}$ was computed using the Overlap (Equation (5)) Vijaymeena and Kavitha (2016), (Equation (6)) Niwattanakul et al. (2013), similarity coefficients using the following formulas.

$$S_{Overlap} = \frac{|(S_{tags}) \cap (R_{tags})|}{\min(|S_{tags}|, |R_{tags}|)}$$ (5)

$$S_{Jaccard} = \frac{|(S_{tags}) \cap (R_{tags})|}{(|S_{tags}| \cup |R_{tags}|)}$$ (6)

### 5.3 Sentence transformer based method

#### 5.3.1 Cross-Lingual sentence transformer based method

Reimers et al.\(^{19}\) have developed the mono-lingual and multi-lingual sentence transformers for sentence embedding Reimers and Gurevych (2019) for a variety of tasks in Natural Language Processing. Initially, they presented Sentence-BERT (SBERT), the modified version of the BERT network using the Siamese and triplet networks which can derive semantically meaningful sentence embedding. SBERT frameworks had been fine-tuned on NLI data\(^{20}\). Authors have evaluated their proposed sentence embedding for a variety of benchmark task and results showed that their proposed sentence embedding outperforms other state-of-the-art sentence embedding methods like the InferSent Conneau et al. (2017), and the Universal Sentence Encoder Devlin et al. (2018). Moreover, these sentence embeddings are extended for more than 100 languages including English, Urdu, Dutch, Chinese, Portuguese, Italian, Spanish, Afrikaans, Albanian, Amharic, Arabic, Armenian, Assamese, Azerbaijani, Basque, Belarusian, Bengali, Georgian, German, Greek, Gujarati, Hausa Hebrew, Hindi, Icelandic, Indonesian, Irish, Italian, Japanese, Javanese, and Korean, etc. These sentence embeddings have been used in a variety of applications including Document Dating Massidda (2020), Objective-Based Hierarchical Clustering

---

\(^{19}\) https://www.sbert.net/, Last visited: 10-2-2021.

\(^{20}\) https://nlp.stanford.edu/projects/snli/, Last visited: 10-2-2021.
Naumov et al. (2020), Generating a Missing Part for Story Completion Mori et al. (2020), Identify Similar Patent Documents Navrozidis and Jansson (2020), Semantic Textual Similarity Guo et al. (2020), Information Retrieval Esteva et al. (2020), and Ranking Clarification Questions Kumar et al. (2020) etc.

The SBERT framework categorizes sentence embeddings into general sentence transformers and special sentence transformers. The general sentence transformers include Bert-base, Bert-large, and bert-base-wikipedia-mean-token etc. The special sentence transformers have been developed for the specific tasks including the Paraphrase Identification, Semantic Textual Similarity, Duplicate Questions Detection, Information Retrieval, and the Bitext Mining Reimers and Gurevych (2020).

To compute the cross-lingual similarity between the cross-lingual text pairs of the proposed corpora, we need a multi-lingual sentence embedding model. Therefore, for this study, we have used the pre-trained Language-agnostic BERT Sentence Embedding (LaBSE) model21 Feng et al. (2020) for the source (English) text and reused (Urdu) text. The main reason to select this model is that it supports 109 languages and works well for cross-lingual text pairs in multiple languages. To compute the similarity between the cross-lingual source-reused text pairs for the English-Urdu language pair, the cross-lingual sentence embedding method was used as follows.

In the first step, English text was converted to the sentence embedding using the LaBSE model. In the second step, Urdu text was converted to the sentence embedding using the LaBSE model. Finally, in the third step, the similarity between the sentence embedding vectors of the English and Urdu texts was computed using the Cosine similarity (see Equ. (3)).

6 Experimental setup

This section describes the corpora, evaluation methodology and the evaluation measures used for the CLTRD experiments applied on the cross-lingual corpora.

6.1 Corpora

Both proposed methods were applied to the proposed CLEU corpora including the Cross-Lingual English Urdu Lexical corpus (CLEU-Lex), Cross-Lingual English Urdu Phrasal corpus (CLEU-Phr), and the Cross-Lingual English Urdu Syntactical corpus (CLEU-Syn). There are a total of 66,485 cross-lingual text pairs (lexical level), out of 22,236 text pairs are WD, 20,315 text pairs are PD, and 23,934 are ND. Similarly, a total of 60,267 cross-lingual text pairs (syntactical level), 20,007 text pairs WD, 16,979 text pairs PD, and 23,281 are ND, and the CLEU-Phr comprises of 60,106 cross-lingual text pairs, out of which 23,862 belong to WD, whereas 15,878 are from PD, and 20,366 belong to ND class at (Phrasal level).

21 The corpus contains 17B mono-lingual sentences and 6B bilingual translation pairs and extracts 768 dimensions averaged vectors of sentence.
6.2 Methods

Firstly, we have applied three cross-lingual word embedding methods: (1) Sum-WE, (2) Average-WE, and (3) a Combination of Sum-WE and Average-WE (called Comb-WE) on the CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora. Secondly, we have applied the Cross-Lingual Semantic Tagger based methods (CLST): (1) CLST using Overlap, (2) CLST using Jaccard, and (3) a Combination of the overlap and Jaccard (called Comb-Both) on the CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora. Thirdly, the combined features of the CLST and the CLWE were passed to the machine learning algorithms (called Comb-ST+WE) for the combined performance. Fourthly, we have applied the cross-lingual sentence transformer method (called CLSTR). Finally, we have applied one more experiment by combining the features of all methods (called comb-all-proposed).

6.3 Evaluation methodology

We have treated the problem of the CLTRD for the English-Urdu language pair as a supervised text classification task for all the three corpora. Two versions of the classification task were used: (1) binary classification task and (2) ternary classification task. The binary classification task aims to distinguish CLTR at two levels: (1) Derived and (2) Non Derived. For this purpose, texts in the Wholly Derived and Partially Derived classes were combined to make a single class, i.e. Derived. The ternary classification task aimed to discriminate the CLTR at three levels: (1) Wholly Derived, (2) Partially Derived, and (3) Non Derived.

For both the binary and ternary classifications tasks, ten different machine learning algorithms were used, including the Bernoulli Naive Bayes (BNB), Gaussian Naive Bayes (GNB), Random Forest (RF), Logistic Regression (LG), Multi-Layer Perception (MLP), Ada Boost (AB), Decision Tree (DT), K-NN, and the Gradient Boosting Classifier (GBC). To better estimate the performance of the machine learning algorithms, a 10-fold cross-validation was used. The similarity/distance scores generated by the various methods (Sect.:5) were used as the input to the machine learning algorithms. Macro-averaged \( F_1 \) scores are reported for all the methods for the binary and ternary classification tasks.

7 Results and analysis

Table 6 and 7 show the weighted average \( F_1 \) scores obtained on the CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora using various CLTRD methods for the binary and ternary classification tasks respectively. In both tables, ‘methods’ refers to the method used for the CLTRD.\(^{23}\)

\(^{22}\) Scikit-learn implementation of these machine learning algorithms was used.

\(^{23}\) For detailed results, see the following link: https://drive.google.com/drive/folders/1-Gqr05n-nBz7419ZuLgpmdzeTz9gd8_r?usp=sharing
For the ternary classification task, overall, for the CLEU-Lex corpus, the best results are obtained using the 'CLST' method ($F_1 = 0.69$ (using Ada Boost (AB), Multi-Layer Perception (MLP), and Gradient Boosting Classifier (GBC)). For the CLEU-Syn corpus, the best results are obtained using the 'Comb-All-proposed' method ($F_1 = 0.82$ (using Random Forest(RF)). Finally, for the CLEU-Phr corpus, the best results are obtained using 'Comb-All-proposed' and the 'CLSTR' methods ($F_1 = 0.78$ (using Ada Boost (AB), Multi-Layer Perception (MLP), Gradient Boosting Classifier (GBC)). Regarding the binary classification task, overall, for the CLEU-Lex corpus, the best results are obtained using the 'CLSTR' method ($F_1 = 0.80$ (using Ada Boost (AB) and Gradient Boosting Classifier (GBC). For the CLEU-Syn corpus, the best results are obtained using the 'Comb-All-proposed' method ($F_1 = 0.92$ (using Ada Boost (AB), Multi-Layer Perception (MLP), Gradient Boosting Classifier (GBC)). Finally, for the CLEU-Phr corpus, the best results are obtained using the 'Comb-All-proposed' and 'CLSTR' methods ($F_1 = 0.94$ (using Ada Boost (AB), Multi-Layer Perception (MLP), Gradient Boosting Classifier (GBC), Gaussian Naive Bayes (GNB), and Decision Tree (DT))). As it can be noted from these results, the performance of the CLTRD methods is different for different corpora. This highlights the fact that the nature of the cross-lingual corpus significantly affects the performance of the CLTRD method. As can be noted
that for the ternary classification, our proposed methods out perform the baseline methods on all the three proposed corpora.

As expected, results for the binary classification task are higher as compared to the ternary classification. This demonstrates the fact that it is easy for a machine learning algorithm to discriminate between the two levels of CLTR (Derived and Non Derived) as compared to the three levels of CLTR (Wholly Derived, Partially Derived, and Non Derived), see Tables (6 and 7). Consequently, there is a large difference in performance between the binary ($F_1 = 0.80$, $F_1 = 0.92$, and $F_1 = 0.94$) and the ternary classification ($F_1 = 0.69$, $F_1 = 0.82$, and $F_1 = 0.78$) tasks for the CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively for all methods used in this study. As can be noted that for the binary classification, our proposed CLWE, and CLSTR methods outperform the baseline methods on all the three proposed corpora. However, the performance of the CLST method is low compared to the baseline approach, which shows that this method is not effective for the CLTRD on our proposed corpora.

For the CLST methods, ‘comb-CLST’ is better than the both (CLST using Overlap, and CLST using Jaccard), the overall performance for the binary classification ($F_1 = 0.54$ using KNN, $F_1 = 0.47$ with all MLA, and $F_1 = 0.53$ with all except

| Methods | CLEU-Lex | CLEU-Phr | CLEU-Syn |
|---------|----------|----------|----------|
| Exp 1.1: CLST using Overlap | 0.24 | 0.23 | 0.25 |
| Exp 1.2: CLST using Jaccard | 0.21 | 0.23 | 0.25 |
| Exp 1.3: Comb-Both | 0.24 | 0.23 | 0.25 |
| Exp 2.1: CLWE-Cosine-Avg | 0.33 | 0.33 | 0.38 |
| Exp 2.2: CLWE-Cosine-Sum | 0.33 | 0.33 | 0.34 |
| Exp 2.3: CLWE-Euclidean-Avg | 0.38 | 0.35 | 0.39 |
| Exp 2.4: CLWE-Euclidean-Sum | 0.42 | 0.37 | 0.43 |
| Exp 2.5: Comb-WE | 0.55 | 0.44 | 0.57 |

| Methods | CLEU-Lex | CLEU-Phr | CLEU-Syn |
|---------|----------|----------|----------|
| Exp 3: Comb-ST+WE | 0.53 | 0.44 | 0.56 |
| Exp 4: CLSTR | 0.69 | 0.78 | 0.78 |
| Exp 5: Comb-All-Proposed | 0.59 | 0.78 | 0.82 |
| Exp 6: Baseline Methods | | | |
| Exp 6.1: Bilingual Dictionary based Method using Overlap | 0.25 | 0.47 | 0.45 |
| Exp 6.2: Bilingual Dictionary based Method using Jaccard | 0.25 | 0.47 | 0.45 |
| Exp 6.3: Combined-Baseline | 0.25 | 0.47 | 0.45 |
GNB) and for the ternary classification \( F_1 = 0.24 \) using KNN, \( F_1 = 0.25 \) with KNN, and \( F_1 = 0.23 \) with all except GNB and KNN) of CLST is still very low for the CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively. The most probable reason is a mostly empty tagset (USAS Urdu tagger did not return most of the tags) for all the CLEU corpora. The possible reason of why the Urdu Semantic Tagger failed to assign Urdu semantic tags to a large number of Urdu words in the CLEU corpora is due to the small size of the Urdu Semantic Lexicons, which contains only 73,189 entries (56,318 are single-word entries and 16,871 are multi-word entries) and the semantically tagged corpus (USA-19 Urdu) contains only 2,213 unique words. The situation gets more complex because the entries in the Urdu semantic lexicon were generated by automatically translating the entries of the English semantic lexicon using the Google and Bing Translators. Out of the 56,318 automatically translated entries, only 2000 were manually inspected and corrected. On the other hand, there is a total of 5955, 3812, and 9302 unique words and 1,09,016, 1,64,375, and 2,17,952 total words in the CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora respectively. As a result, the majority of words in the CLEU corpora are not found in the Urdu Semantic Lexicon, and consequently, no Urdu semantic tag is assigned to them. Consequently, the majority of similarity scores for cross-lingual text pairs in the CLEU corpora are 0, which makes it difficult for the machine learning algorithms to discriminate between various levels of the CLTRD.

For the CLWE methods, the best results are obtained using the ‘Comb-WE’ method (for the binary classification \( F_1 = 0.64 \) using DT, \( F_1 = 0.66 \) using RF, and \( F_1 = 0.61 \) using DT) for the CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively and (for the ternary classification \( F_1 = 0.55 \) using DT & RF, \( F_1 = 0.57 \) DT and \( F_1 = 0.44 \) using DT & RF) for the CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively. As can be noted that these results are better than the CLST, and highlight the fact that CLWE methods are effective in detecting the CLTR if WE models are trained on large data. The possible reason is, as both (the Google pre-trained word embedding model and MK-PUCIT) are trained on large data, therefore both pre-trained models can extract high quality feature vectors for the cross-lingual pairs.

The combination of CLWE and CLST (i.e., ‘Comb-ST+WE’) was passed to MLA, for ‘Comb-ST+WE’, results are obtained for the binary \( F_1 = 0.62 \) using RF, \( F_1 = 0.65 \) using DT, and \( F_1 = 0.61 \) using DT & RF), and the ternary classification \( F_1 = 0.53 \) using DT, \( F_1 = 0.56 \) using RF, and \( F_1 = 0.44 \) using the DT & RF) tasks. This shows that the performance of comb-all-proposed is slightly lower for the CLEU-Lex for both the binary and the ternary classification tasks. However, for the CLEU-Syn, binary scores are the same as ‘Comb-WE’ and slightly lower for the the ternary classification and the performance for the CLEU-Phr is the same as ‘Comb-WE’ for the both classification tasks. This shows that this combination of features is not helpful to improve performance.

For the CLSTR method, the best results are obtained (for the binary classification \( F_1 = 0.80 \) using AB and GBC, \( F_1 = 0.92 \) using AB, MLP, GBC, and GNB, and \( F_1 = 0.94 \) AB, MLP, GBC, GNB, and DT) for the CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively and (for ternary classification \( F_1 = 0.69 \) using AB, MLP and GBC, \( F_1 = 0.78 \) using RF and \( F_1 = 0.78 \) using AB, GBC and MLP) for the
CLEU-Lex, CLEU-Syn, and the CLEU-Phr respectively. As can be noted that these results are better than all the applied methods, highlighting the fact that the CLSTR methods are more effective in detecting the CLTR if the CLSTR models are trained on large data. The possible reason is, that the LaBSE is trained on 6 billion multi-lingual pairs and specially designed for the translated sentence pairs in two languages, therefore it can extract the best quality feature vectors for the cross-lingual pairs. However, the results for the CLEU-Lex is still much lower than the other two corpora for the ternary classification task. The possible clarification for lower performance is that the LaBSE is trained on the sentence data and can work better for finding matching chunks of longer and different lengths.

The combination of CLWE, CLST and CLSTR (i.e., 'comb-all-proposed') were passed to the MLA, for Combml-comb-all-proposed, results are obtained for binary ($F_1 = 0.71$ using GNB, AB, LR, and GNB, $F_1 = 0.92$ using GNB, LG, MLP, and GBC, and $F_1 = 0.94$ using LR, AB, MLP, KNN, and GBC), and ternary classification ($F_1 = 0.59$ using GBC, $F_1 = 0.82$ using RF and $F_1 = 0.78$ using the AB, MLP, and GBC) tasks. This shows the performance of comb-all-proposed is slightly lower for the CLEU-Lex for both the binary and ternary classification tasks and for the CLEU-Syn binary classification task performance is the same as that of the CLSTR, and it increases from 0.78 to 0.82 for the ternary classification task. However, for the CLEU-Phr, the binary and ternary scores are the same as the 'CLSTR' method. This shows that this combination of features is not helpful to improve the performance except the CLEU-Syn for the ternary classification task.

The difference in the performance of the baseline methods and proposed (CLWE, CLST, and CLSTR) methods is quite large. The possible reason for the low performance of baseline methods is that, in a machine-readable bi-lingual Dictionary, a word may have multiple translations. When we search for the translation of a word (with multiple possible translations), the machine-readable Bilingual Dictionary may not return the most appropriate translation of that word. Besides this, the size of the dictionary is 9822. On the other hand, there is a total of 5955, 3812, and 9302 unique words and 1,09,016, 1,64,375, and 2,17,952 total words in the CLEU-Lex, CLEU-Syn, and the CLEU-Phr corpora respectively. As a result, the majority of words in the CLEU corpora are not found in the dictionary, and consequently, no translation is assigned to them. Consequently, the majority of similarity scores for the cross-lingual text pairs in the CLEU corpora are 0, which makes it difficult for the machine learning algorithms to discriminate between the various levels of the CLTR.

Among the machine learning algorithms, different algorithms show different behavior for different methods for the binary and ternary classification tasks. Mostly, the Gradient Boosting Classifier, Ada Boost, Multi-Layer Perception, and in few cases, KNN and Gradient Boosting Classifier, and Ada Boost mostly occurred for the binary classification for all corpora. The Ada Boost, Multi-Layer Perception, and the Gradient Boosting Classifier showed the best performance for the ternary classification for the CLEU-Lex and CLEU-Phr, except for the RF shows the best performance for 'comb-all-proposed' for the CLEU-Lex ternary classification task only. The performance deviates in different experiments, so there is no universal and conclusive machine learning algorithm that can perform best for the corpora.
To conclude, the main findings obtained after our extensive experimentation are as follows. Firstly, there is a significant difference in performance for the binary and ternary classification tasks, which shows that it becomes more difficult to discriminate the three levels of the CLTR as compared to two levels of text reuse. Secondly, different CLTRD methods show different performance on different CLTRD corpora at the lexical, syntactical, and phrasal levels. This shows that the length of text in a cross-lingual text reuse corpus affects the performance of the CLTRD method. Thirdly, different algorithms show different behavior for different methods for the binary and ternary classification tasks. Fourthly, the SBERT framework extracts more qualitative feature vectors for the phrasal level as compared to the lexical level.

8 Conclusion

The rapid increase in online multi-lingual content has gained the interest of the research community of CLTRD. A major drawback is the unavailability of gold-standard benchmark corpora for the development of CLTRD methods, especially for the language (e.g., Urdu), which is highly under-resourced in general. In this paper, we address this gap by presenting the gold-standard benchmark corpora at the lexical, syntactical, and the phrasal levels for the English-Urdu language pair with simulated examples of reuse for the first time. This study presents three large benchmark corpora for the CLTRD for English-Urdu language pair at the lexical, syntactical, and the phrasal levels. All three proposed corpora are manually annotated at three levels of rewrite including Wholly Derived, Partially Derived, and Non Derived. We have also applied the CLWE, CLST, CLSTR methods on our proposed corpora. Results show that for the binary classification, the best results on the CLEU-Lex corpus were obtained using the cross-lingual sentence transformer ($F_1 = 0.80$). For the CLEU-Syn and CLEU-Phr corpora, the best results were obtained using the cross-lingual sentence transformer and the combination of CLWE, CLST, and CLSTR methods ($F_1 = 0.92$ on CLEU-Syn and $F_1 = 0.94$ on CLEU-Phr). For the ternary classification, the best results on the CLEU-Lex corpus were obtained using the cross-lingual sentence transformer method ($F_1 = 0.69$). For the CLEU-Syn corpus, the best results were obtained using a combination of the CLWE, CLST, and CLSTR methods ($F_1 = 0.82$). For the CLEU-Phr corpus, the best results were obtained using the cross-lingual sentence transformer and a combination of CLWE, CLST, and CLST methods ($F_1 = 0.78$). In the future, we plan to employ Kullback-Leibler distance at the cross-lingual level to reduce the text reuse detection search space.

References

Abdi, A., Idris, N., & Alguliyev, R. M. (2015). PDLK: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications, 42*(22), 8936–8946.

Alaa, Z., Tian, S., & Abdulameer, M. (2016). Cross-language plagiarism of Arabic-English using linear logistic regression. *Journal of Theoretical & Applied Information Technology, 83*(1), 23.
Develop corpora and methods for cross-lingual text reuse…

Alfikri, Z. F., & Purwarianti, A. (2012). The construction of Indonesian-English cross language plagiarism detection system using fingerprinting technique. *Jurnal Ilmu Komputer dan Informasi*, 5(1), 16–23.

Aljohani, A., & Mohd, M. (2014). Arabic-English cross-language plagiarism detection using winnowing algorithm. *Information Technology Journal*, 13(14), 2349.

Al-Suhaili, M., Hazaa, M. A., & Albared, M. (2018). Arabic English cross-lingual plagiarism detection based on keyphrases extraction, mono-lingual and machine learning method. *Asian Journal of Research in Computer Science*, 1-12.

Asghari, H., Khoshnava, K., Fatemi, O., & Faili, H. (2015). Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. *Notebook for PAN at CLEF*.

Bakhteev, O., Ogaltsov, A., Khazov, A., Safin, K., & Kuznetsova, R. (2019, September). CrossLang: the system of cross-lingual plagiarism detection. In *Workshop on Document Intelligence at NeuIPS 2019*.

Barrón-Cedeño, A., Rosso, P., Agirre, E., & Labaka, G. "Plagiarism detection across distant language pairs." *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010.

Barrón-Cedeño, A., Rosso, P., Devi, S. L., Clough, P., & Stevenson, M. (2013). Pan@ 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, Heidelberg.

Barrón-Cedeño, Alberto, Gupta, Parth, & Rosso, Paolo. (2013). Methods for cross-language plagiarism detection. *Knowledge-Based Systems*, 50, 211–217.

Ceska, Z., Toman, M., & Jezek, K. (2008). Multilingual plagiarism detection. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (pp. 83-92). Springer, Berlin, Heidelberg.

Chang, Chia-Ming., Chang, Chia-Hsuan., & Hwang, San-Yih. (2020). Employing word mover’s distance for cross-lingual plagiarized text detection. *Proceedings of the Association for Information Science and Technology*, 57(1), e229.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *ArXiv Preprint arXiv:1705.02364*.

Daille, B., & Morin, E. "Effective compositional model for lexical alignment." *IJCNLP 2008: Third International Joint Conference on Natural Language Processing*. Vol. 1. 2008.

de Souza, J. V. A., Oliveira, L. E. S. E., Gumiel, Y. B., Carvalho, D. R., & Moro, C. M. C. (2020, March). Exploiting Siamese neural networks on short text similarity tasks for multiple domains and languages. In *International Conference on Computational Processing of the Portuguese Language* (pp. 357-367). Springer, Cham.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D., & Socher, R. (2020). Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *ArXiv Preprint arXiv:2006.09595*.

FA14-MSCS, I. M. Measuring cross-lingual text reuse at sentence/passage Level. Diss. 2016.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

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).

Ferrero, J., Agnes, F., Besacier, L., & Schwab, D. (2017). Using word embedding for cross-language plagiarism detection. *ArXiv Preprint arXiv:1702.03082*.

Flores Sáez, E., Barrón-Cedeño, L. A., Moreno Boronat, L. A., & Rosso, P. (2015). Cross-language source code re-use detection using latent semantic analysis. *Journal of Universal Computer Science*, 21(13), 1708–1725.

Forner, P., Karlgren, J., Womser-hacker, Ch., Potthast, M., Gollub, T., Hagen, M., Graegger, J., Kiesel, J., Michel, M., Oberländer, A., Barrón-cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th international competition on plagiarism detection. In: Forner, P., Karlgren, J., Womser-Hacke, C. (eds.) *Notebook Papers of CLEF 2012 LABs and Workshops, CLEF-2012 17–20 September. Rome, Italy*. Franco-Salvador, M., Gupta, P., & Rosso, P. (2013, March). Cross-language plagiarism detection using a multilingual semantic network. In *European Conference on Information Retrieval* (pp. 710-713). Springer, Berlin, Heidelberg.
Franco-Salvador, M., Gupta, P., Rosso, P., & Banchs, R. E. (2016). Cross-language plagiarism detection over continuous-space and knowledge graph-based representations of language. *Knowledge-based Systems, 111*, 87–99.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013, June). PPDB: The paraphrase database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 758-764).

Ghannay, S., Favre, B., Esteve, Y., & Camelin, N. (2016). Word embedding evaluation and combination. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16) (pp. 300-305).

Grefenstette, G. (1998). *The problem of cross-language information retrieval* (pp. 1–9). Boston, MA: Cross-language information retrieval. Springer.

Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A., & Abd-Almageed, W. (2020). Cord19sts: Covid-19 semantic textual similarity dataset. ArXiv Preprint arXiv:2007.02461.

Gupta, Parth, Alberto Barrón-Cedeno, and Paolo Rosso. "Cross-language high similarity search using a conceptual thesaurus." International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, Berlin, Heidelberg, 2012.

Hadgu, A. T. (2018). Cross-lingual Short-text Matching with Deep Learning. ArXiv Preprint arXiv:1811.05569.

Haneef, I., Adeel Nawab, R. M., 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.

Hanif, I., Nawab, R. M. A., Arbab, A., Jamshed, H., Riaz, S., & Munir, E. U. (2015). Cross-language Urdu–English (clue) text alignment corpus. Working notes papers of the CLEF.

Kanwal, S., Malik, K., Shahzad, K., Aslam, F., & Nawaz, Z. (2019). Urdu named entity recognition: Corpus generation and deep learning applications. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1), 1–13.

Kent, C. K., & Salim, N. (2010, September). Web based cross language plagiarism detection. In 2010 Second International Conference on Computational Intelligence, Modelling and Simulation (pp. 199-204). IEEE.

Kenter, T., & De Rijke, M. (2015). Short text similarity with word embeddings. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1411-1420).

Khorsh, A., Cherroun, H., & Schwab, D. (2018). A two-level plagiarism detection system for Arabic documents. *Cybernetics and Information Technologies.*, 18(1), 1003.

Kothwal, R., & Varma, V. (2013). 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, Heidelberg.

Kumar, V., Raunak, V., & Callan, J. (2020). Ranking clarification questions via Natural Language Inference. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 2093-2096).

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. In 2016 4th International Conference on Cyber and IT Service Management (pp. 1-6). IEEE.

Li, X., Chen, M., & Zeng, Z. (2018, October). Cross-Lingual semantic textual similarity modeling using neural networks. In China Workshop on Machine Translation (pp. 52-62). Springer, Singapore.

Massidda, R. (2020). rmassidda@ DaDoEval: Document dating Using sentence embeddings at EVALITA 2020. In Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian: Final Workshop (EVALITA 2020), Online. CEUR.org.

Mikolov, T., Yih, W. T., & Zweig, G. (2013, June). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 746-751).

Mori, Y., Yamane, H., Mukuta, Y., & Harada, T. (2020, December). Finding and generating a missing part for story completion. In Proceedings of the The 4th Joint Sighum Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (pp. 156-166).

Mueller, J., & Thyagarajan, A. (2016). Siamese recurrent architectures for learning sentence similarity. In thirtieth AAAI conference on artificial intelligence.

Muhammad, S. (2020). Mono-and cross-lingual paraphrased text reuse and extrinsic plagiarism detection (Doctoral dissertation, Lancaster University).
Muneer, I., & Nawab, R. M. A. (2021). Cross-lingual text reuse detection using translation plus monolingual analysis for English-Urdu language pair. *Transactions on Asian and Low-Resource Language Information Processing*, 21(2), 1–18.

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.

Napoles, C., Callison-Burch, C., & Post, M. (2016, June). Sentential paraphrasing as black-box machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 62-66).

Naumov, S., Yaroslavtsev, G., & Avdiukhin, D. (2020). Objective-Based hierarchical clustering of deep embedding vectors. arXiv preprint arXiv:2012.08466.

Navrozidis, J., & Jansson, H. (2020). Using Natural Language Processing to identify similar patent documents. LU-CS-EX.

Neculoiu, P., Versteegh, M., & Rotaru, M. (2016). Learning text similarity with siamese recurrent networks. In Proceedings of the 1st Workshop on Representation Learning for NLP (pp. 148-157).

Nicosia, M., & Moschitti, A. (2017, November). Accurate sentence matching with hybrid Siamese networks. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2235-2238).

Niawattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, No. 6, pp. 380-384).

Ozsoy, M. G. (2016). From word embeddings to item recommendation. ArXiv Preprint arXiv:1601.01356.

Pavlick, E., & Callison-Burch, C. (2016, August). Simple PPDB: A paraphrase database for simplification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 143-148).

Pei, W., Tax, D. M., & van der Maaten, L. (2016). Modeling time series similarity with Siamese recurrent networks. ArXiv Preprint arXiv:1603.04713.

Pelevina, M., Arefyev, N., Biemann, C., & Panchenko, A. (2017). Making sense of word embeddings. ArXiv Preprint arXiv:1708.03390.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).

Piao, S. S., Bianchi, F., Dayrell, C., D’egidio, A., & Rayson, P. (2015). Development of the multilingual semantic annotation system. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1268-1274).

Piao, S. S., Rayson, P., Archer, D., & McEnery, T. (2005). Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language*, 19(4), 378–397.

Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., & Rosso, P. (2009). A statistical approach to cross-lingual natural language tasks. *Journal of Algorithms*, 64(1), 51–60.

Potthast, M., Eiselt, A., Barrón-Cedeño, L. A., Stein, B., & Rosso, P. (2011). Overview of the 3rd international competition on plagiarism detection. In CEUR workshop proceedings (Vol. 1177). CEUR Workshop Proceedings.

Potthast, M., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1), 45–62.

Ranasinghe, T., Orasan, C., & Mitkov, R. (2019). Semantic textual similarity with siamese neural networks. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019) (pp. 1004-1011).

Rayson, P., Archer, D., Piao, S. L., & McEnery, T. (2004). *The UCREL Semantic Analysis System*. Workshop Beyond Named Entity Recognition Semantic Labelling for NLP Tasks: Proc.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. ArXiv Preprint arXiv:1908.10084.

Reimers, N., & Gurevych, I. (2020). Making mono-lingual sentence embeddings multilingual using Knowledge Distillation. ArXiv preprint arXiv:2004.09813.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. ArXiv Preprint arXiv:1706.05098.

Sameen, S., Sharjeel, M., Nawab, R. M. A., Rayson, P., & Muneer, I. (2017). Measuring short text reuse for the Urdu language. *IEEE Access*, 6, 7412–7421.
Scanlon, P. M., & Neumann, D. R. (2002). Internet plagiarism among college students. *Journal of College Student Development, 43*(3), 374–385.

Shafi, J. (2019). An Urdu semantic tagger-lexicons, corpora, methods and tools (Doctoral dissertation, Lancaster University).

Shi, H., Wang, C., & Sakai, T. (2020). A Siamese CNN architecture for learning Chinese sentence similarity. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop (pp. 24-29).

Štajner, T., & Mladenić, D. (2019). Cross-lingual document similarity estimation and dictionary generation with comparable corpora. *Knowledge and Information Systems, 58*(3), 729–743.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. “Learning sentiment-specific word embedding for twitter sentiment classification.” *ACL (1)*. 2014.

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. ArXiv Preprint arXiv:1604.00425.

Varior, R. R., Shuai, B., Lu, J., Xu, D., & Wang, G. (2016). A Siamese long short-term memory architecture for human re-identification. In European conference on computer vision (pp. 135-153). Springer, Cham.

Vijaymeena, M. K., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal, 3*(2), 19–28.

Vinayakumar, R., & Soman, K. P. (2020). Siamese neural network architecture for homoglyph attacks detection. *ICT Express, 6*(1), 16–19.

Wang, J., Qin, Y., Peng, Z., & Lee, T. (2019). Child speech disorder detection with Siamese Recurrent network using speech attribute Features. In Interspeech (pp. 3885-3889).

Wieting, J., Bansal, M., Gimpel, K., & Livescu, K. (2015). From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics, 3*, 345–358.

Xu, X., Ma, B., Chang, H., & Chen, X. (2017). Siamese recurrent architecture for visual tracking. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 1152-1156). IEEE.

Yang, J., Zou, H., Zhou, Y., & Xie, L. (2019). Learning gestures from wifi: A Siamese recurrent convolutional architecture. *IEEE Internet of Things Journal, 6*(6), 10763–10772.

Zhang, L., & Moldovan, D. (2018). Rule-based vs. neural net methods to semantic textual similarity. In Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing (pp. 12-17).

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

**Authors and Affiliations**

Iqra Muneer1,2 · Rao Muhammad Adeel Nawab1

✉️ Iqra Muneer  
fa18-pcs-002@cuilahore.edu.pk; iqramuneer@uet.edu.pk  
Rao Muhammad Adeel Nawab  
adeelnawab@cuilahore.edu.pk

1 COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan  
2 University of Engineering & Technology Lahore, Narowal Campus, Pakistan