# DOMAIN-SPECIFIC CROSS-LINGUAL URDU TO ENGLISH (CLUE) PLAGIARISM DETECTION

F. Shahzad, S. Jabeen, M. Pasha[*], B. Majeed[*], and X. Gao[**]

Department of Computer Science, Pakistan Institute of Engineering and Technology, 60000 Multan
[*]Department of Information Technology, Bahauddin Zakariya University, 60000 Multan
[**]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
Corresponding author's email: farrukhshahzad@piet.edu.pk

**ABSTRACT:** Plagiarism is an act of copying someone's text without reference. Cross-lingual plagiarism detection (CLPD) deals with discovering and retrieving of the copied words and sentences in a bilingual scenario. There have been various attempts to detect the cross-lingual plagiarism in settings like English-to-German, English-to-Spanish, and Arabic-to-English. However, no system or framework is available for Urdu-English CLDP. This paper presents a new framework for detection of Urdu-English plagiarism using *Translate plus Mono Lingual Analysis* technique. The framework used CLUE (Cross-lingual Urdu-English) corpus which contains documents from two domains: computer literature; and general area. The main outcome of paper is to detect the Cross lingual plagiarism (Urdu to English). Translational quality of Google and Bing is analyzed for translation of source documents. Empirical results have shown that plagiarism ratio varied with translational tools for different plagiarism cases. Experiments have shown that palgiarism ratio in higher in Near Copy documents, moderate in light Revision and least in Heavy Revision documents. This research is useful to understand the scenarios in which certain translational tools are effective.

**Key words:** Cross-lingual Plagiarism, Plagiarism detection, Text similarity analysis, n-Grams, Translational tools.

*(Received 17-10-2017 Accepted 22-06-2018)*

## INTRODUCTION

The word plagiarism is essentially derived from a Latin word 'plagiare' which means stealing something. Plagiarism, an act of stealing someone else's actual words or ideas as one's own, is usually seen as the legitimate offense (Chiu, 2010; Gupta, 2012). It is a way of reusing writings, plans and thoughts of someone or a group by an unauthorized person, without author's permission (Vinod, 2011).

Plagiarism detection refers to detecting the copied text that is used without original author's permission or without quoting proper references (Ferrero, 2017; Luke, 2014). Plagiarism detection is the spontaneous detection of copied text and finding the original sources (Asghari, 2015; Potthast, 2011; Potthast, 2013). There is no system available to detection Urdu to English plagiarism. Urdu language is totally different from other languages in terms of grammar, vocabulary, sentences and flow. Corpus was required to make the experiments to check the accuracy of plagiarism detection. Translation tools were used to translate the Urdu documents into the English document.

It is almost impossible to manually analyze the huge amount of text in millions of electronic documents to identify the original source of copied sentences. Due to this reason, computational knowledge of plagiarism detection has now become an eminent research direction. Plagiarism detection is mainly divided into two main groups: Intrinsic Plagiarism Detection (Oberreuter, 2011; Seaward, 2015) reviewing that the complete text is written by the same author; and Extrinsic Plagiarism Detection (Alzahrani, 2010; Asghari, 2015) finding the original sources of plagiarized fragments (Hanif, 2015). Extrinsic Plagiarism Detection can be categorized further as (1) Mono-lingual Plagiarism Detection (Gupta, 2011) and (2) Cross-lingual Plagiarism Detection. Intrinsic Plagiarism detection is not discussed in this paper since it is out of the scope of the current research work.

Cross-lingual plagiarism detection (CLPD) deals with an automatic resemblance discovering and retrieving the copied words and sentences in a bilingual scenario (Barrón-Cedeño, 2013b; Kent, 2010; Potthast, 2013; Omar, 2013; Vinod, 2011). There are different techniques to detect plagiarized text in the cross-lingual scenario. E.g., calculating the syntax-based similarity between different texts (Ali, 2011).

There are various extents to which a text can be plagiarized. These can be generalized into three cases of plagiarism (Hanif, 2015): (i) Near Copy (NC), some portion of an article or file is directly plagiarized from the original source file, but no reference is provided. (ii) Light Revision (LR), some part of the article is plagiarized from the original source by making some modifications in the content using any online rewriting tool, but the structure and style of fragments remain

same. (iii) Heavy revision, the plagiarized portion is rephrased by a professional writer.

Character Dot-Plot (Gopalakrishnan, 2013; Rabiu, 2014) is a statistical plot chart containing the character-based data points designed on a simple scale, usually using highlighted dots to analyze the resemblances in various texts. Dot-plot (Park, 2015) is a method for envisioning patterns of words or character matches in hundreds of lines and code. Patterns can be discovered manually or detected by an automatic tool. Patterns are inferred through a language of visualization.

An arrangement of words is tokenized and then plotted from left to right and top to bottom by plotting a dot where tokens are matched. The main diagonal of dots shows the exact token matching (Ekbal, 2012; Field, 2013; Jänicke, 2014).

Cross-Lingual Character N-Grams (CL-CNG) (Franco-Salvador, 2013) is composition and syntax-based model. This model uses the n-gram characters-based technique (Barrón-Cedeno, 2010; Potthast, 2011). Normally 3-grams or 4-grams (Ahmadzia, 2015; Markov, 2017; Sapkota, 2015) characters are used to find similarities in the source and suspicious files.

CL-CNG technique was first introduced in 2004. This technique performs efficiently only for the dialects sharing the lexical and syntactic resemblances (Indo-European families). Words N-gram (Buntinx, 2015; Coffee, 2014; Forstall, 2014) technique is also used to check the plagiarism in the same manner.

Cross Lingual Alignment based Similarity Analysis (CL-ASA) is a parallel corpora-based model (Franco-Salvador, 2013). It is based on the statistical automatic translation expertise (Mitkov, 2016). CL-ASA (Franco-Salvador, 2014; Flores, 2015) does well with the professional and automatic paraphrases (because of the nature of corpus used), (Danilova, 2013).

Cross Lingual Latent Semantic Indexing (CL-LSI) is also a parallel corpora based model (Potthast, 2011). It is a corporate scheme applied in IR systems for the term-document association. CL-LSI performs the concealed semantic indexing (Evangelopoulos, 2013; Shen, 2014).

Cross Lingual Kernel Canonical Correlation Analysis (CL-KCCA) performs better than the LSI on the same datasets, although it is based on SVD as well (Nadil, 2016; Souilem, 2017). CL-KCCA does a kernel acknowledged correlation analysis. It is parallel corpora based model (Xiao, 2014).

Translate plus Mono Lingual Analysis (T+MA) is another method of cross-lingual similarity analysis model. In this approach, both source and suspicious files are translated in the same language first and then plagiarism detection technique is applied (Barrón-Cedeño, 2013a). T+MA is an expensive method regarding computations, and there still is the absence of effective automatic translators for various language pairs (Danilova, 2013).

## MATERIALS AND METHODS

Different corpuses were available in different languages in the cross-lingual domain. CLUE (Cross-lingual Urdu English) corpus (Hanif, 2015) was used in this paper to detect the plagiarism because this corpus has multiple levels of plagiarized cases in Urdu and English. A framework is proposed in this paper to detect the plagiarism in cross lingual context by empirical study of various cross lingual detection patterns (Ferrero, 2017; Omer, 2013). Two main phases existed in the proposed framework: Translation and plagiarism detection. First, all selected files from both domains were translated into English files then a monolingual technique was used to detect plagiarism. N-Gram technique was used to check the plagiarized fragments. Bing and Google Translator were used to translate the documents from Urdu to English before checking the plagiarism. Randomly any suspicious Urdu document from the CLUE Corpus can be chosen to check the plagiarism by this framework. The proposed architecture is shown in Fig. 1.

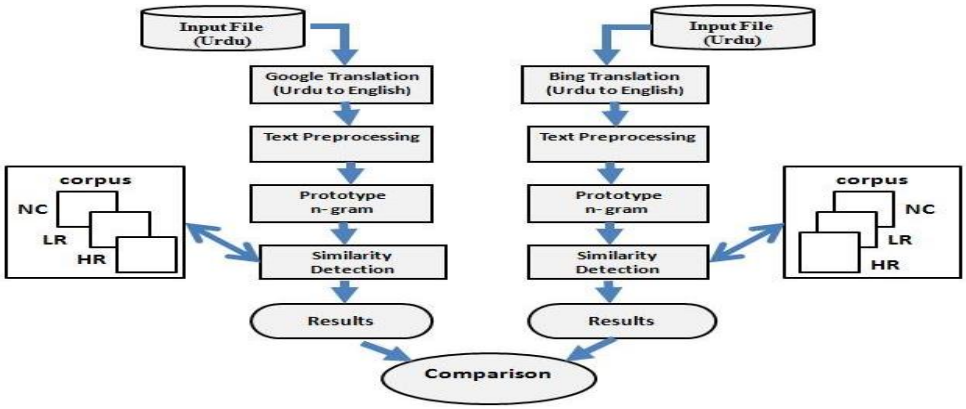


**Figure-1: Proposed Architecture for CLPD**

**Various stages of the proposed framework are explained below:** In translation phase, suspicious Urdu files from CLUE corpus were translated into English. For this purpose, two different translators were used: Google translator and Bing Translator. This phase was valuable for analyzing the impact of using various translational tools on the overall performance of T+MA based cross-lingual plagiarism detection approach.

Text processing phase involved various steps to preprocess the raw text of input files and prepare it for similarity detection. After translation from Urdu to English language, now both files (source and suspicious) were in the same language. Hence, text preprocessing was applied to both sets of documents in the first part of this phase. Text preprocessing contained stopping word removal, special characters removal and stemming (removal of derivational affixes).

The input text was then tokenized into 4-grams using N-grams technique. Four consecutive words were considered as a single token. These tokens were used to represent the suspicious file. If these tokens match with the tokens of the source documents, then these were considered as plagiarized fragments.

Similarity detection was done by comparing the plagiarized fragments in a suspicious file against the tokens of the source file using exact match of N-grams (4-grams in the case of the proposed approach) technique. Similarity detection of the prototype (developed based on the proposed framework) is shown in Fig. 2. Code of algorithm for comparing tokens and calculating plagiarism in percentage are shown in Fig-3 and Fig-4.

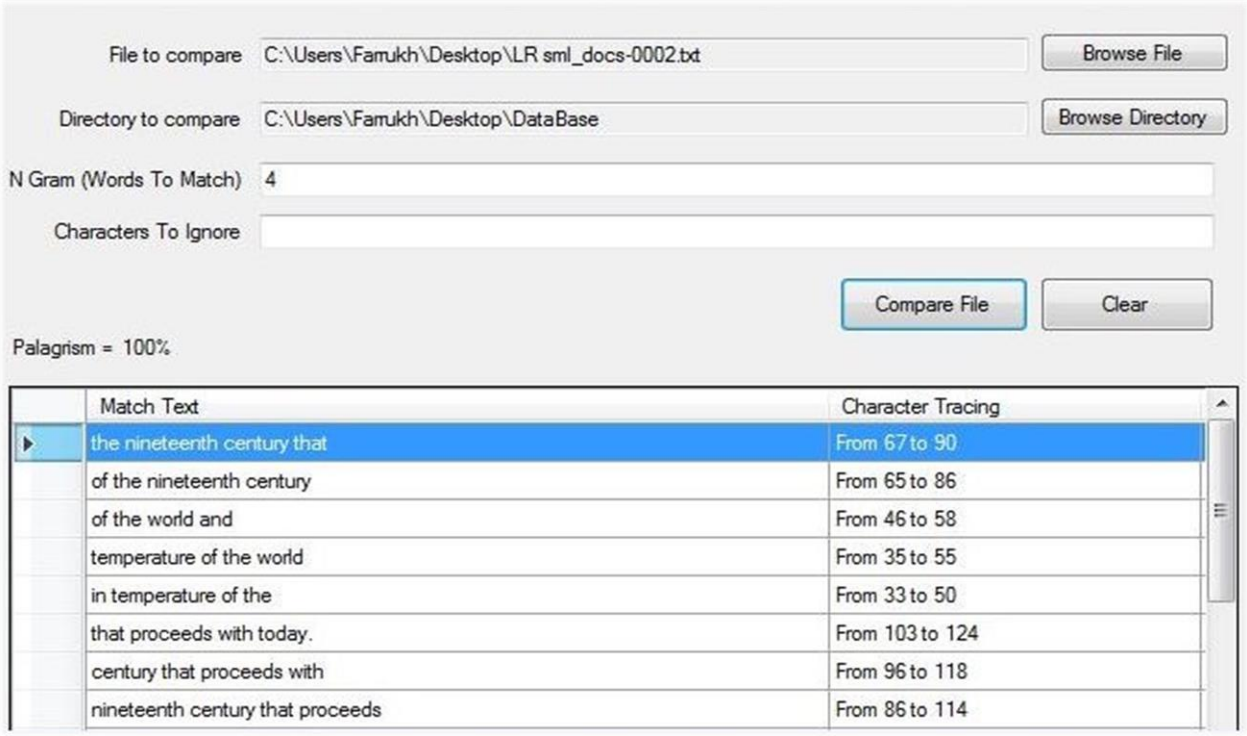


**Figure-2: Similarity detection in software using word 4-grams tokenization.**

```
foreach (string text in SourceFileContent)
{
    if (fileContent.Replace(" ", "").ToLower().Trim().Contains(text.Replace(" ", "").ToLower().Trim()))
    {
        CompareResult compareResult = new CompareResult();
        compareResult.matchText = text;
        compareResult.filePath = filesList[i];
        // set value of chracter tracing
        compareResult.charTracing = SetCharTracingText(fileContent, text);

        CompareResultBag.Add(compareResult);
    }
}
```

**Figure-3: Code for Comparing Strings**

```
// set plagiarism result in percentage
private void SetPlagiarismPercentage()
{
    try
    {
        int percentage = 0;

        var sourceFileList = SourceFileContent.Distinct().ToList();
        var compareResultsList = CompareResultBag.GroupBy(item => item.matchText)
                                                 .Select(grp => grp.OrderBy(item => item.matchText).First())
                                                 .ToList();

        if (sourceFileList.Count > 0 && compareResultsList.Count > 0)
            percentage = (int)(Convert.ToDouble((double)compareResultsList.Count / (double)sourceFileList.Count) * 100);

        labelPalagrismPer.Visible = true;
        labelPalagrismResult.Visible = true;

        labelPalagrismResult.Text = percentage.ToString() + "%";
    }
    catch (Exception ex)
    {
        throw ex;
    }
}
```

**Figure-4: Code for Calculating Plagiarism in Percentage**

In the last phase, the plagiarism results of both domains were compared to analyze the performance difference of proposed framework using two different translational services.

## RESULTS AND DISCUSSION

The proposed approach has used CLUE (Cross-lingual Urdu-English) corpus. A prototype was built to detect plagiarism in a cross-lingual manner. T+MA technique was used to develop this prototype. Monolingual plagiarism detection technique used in the prototype is words n-gram. For experiments, arbitrarily 10 files were selected from the documents having large text, 10 having moderate and 10 having minor text from Computer Science portion and similarly from General Topics' portion. Bing and Google translating services were used to convert Urdu files into English. Total documents translated from Google and Bing translator were 60. Plagiarism ratio was used to measure and compare the performance of various experimental setups for cross-lingual plagiarism detection methods (Asghari, 2016). A summary of attributes of source and suspicious fragments of CLUE corpus is depicted in Table 1.

**Table-1: Attributes of the source and suspicious fragments of CLUE corpus.**

| Level of fragments (words) | Level name | Total no. of fragments (270) | |
|---|---|---|---|
| | | CS(180) | General (90) |
| <=50 | Sentence (small) | 100 | 50 |
| >50 and <=100 | Paragraph (medium) | 50 | 25 |
| >=100 and <=200 | Essay (large) | 30 | 15 |

**General topics domain Vs. CS Literature domain:** Performance comparison of the proposed system in Computer Science domain and General topics domain using both translators is shown in Fig. 5. Plagiarism ratio in General topics domain was higher in all plagiarism cases as compared to that of the CS domain. Most plagiarism was detected in NC, then LR and least in HR in CLUE Corpus (Hanif, 2015; Rana, 2016; Vani, 2017). In results, the sum of the plagiarized ratio was not equal to 100% because not all the text of the suspicious document was plagiarized.

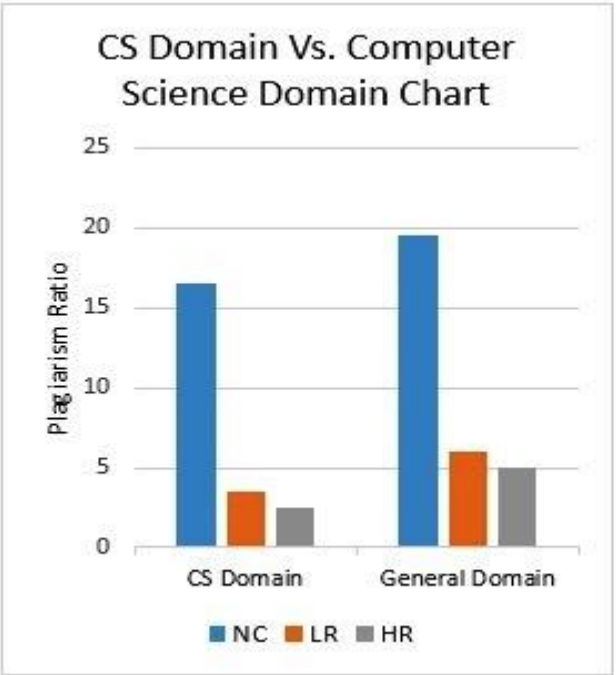


**Fig. 5. Results Comparison between CS Domain and General Topics Domain using both Translators**

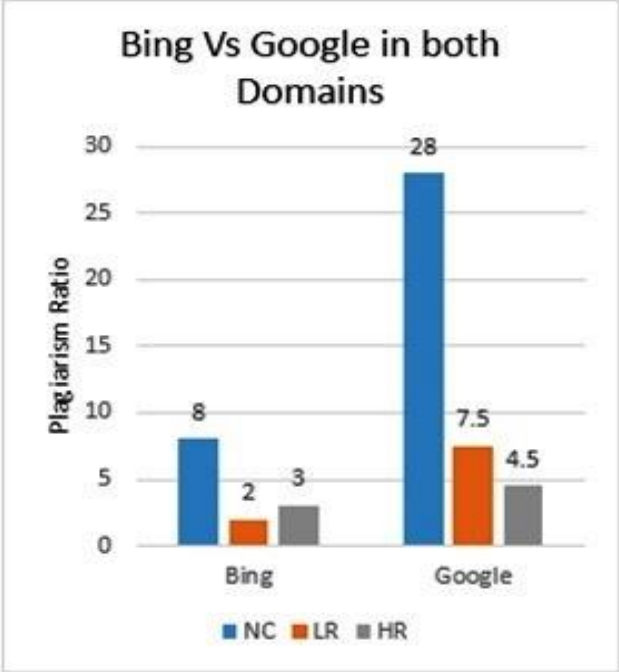


**Fig. 6. Results Comparison between results of Bing Translator and Google Translator**

In the Computer Science domain, plagiarism ratio of NC case has the highest value as shown in (Chong, 2010), and HR case has lowest plagiarized ratio using Bing and Google translators (Saffari, 2017). Same as the previous results, in the case of the general domain, plagiarism ratio of NC has the highest value using both translators (Mahmoodi, 2014).

A comparison of detected plagiarism ratio in both domains using Google and Bing translators is shown in Fig. 6. Results indicate that the files after translating from the Google translator show higher plagiarism ratio. This distinction was potential because of the nature of selected corpus.

Results indicate that the Plagiarism ratio was distinct when files were translated from two online translational tools before detecting plagiarism. Overall, 8% copied text lied in Near Copy (NC), 2% lied in Light Revision (LR), and 3% lied in Heavy Revision (HR) when all files were translated using Bing Translator. However, when files of both domains were translated from Google translator, 28% copied text lied in Near Copy (NC), 7.5% lied in Light Revision (LR), and 4.5% lied in the Heavy Revision (HR) category. Plagiarism ratio was highest in NC as shown in (Chong, 2010; Stapleton, 2012), moderate in LR and least in HR in both domains (Clough, 2011).

**Conclusion:** This paper presented a T+MA-based framework for cross-lingual plagiarism detection using two translational tools: Bing and Google. Similarity comparison of source and suspicious files was made by using N-grams-based similarity matching scheme. Experimental results have shown the different ratio of plagiarism against different plagiarism cases (NC, LR, and HR) in the corpus. Comparative study of percentile ratio amongst these levels showed that HR has the least plagiarism ratio (Cheema, 2015) and the plagiarism in the Google translated files was higher than the Bing translated files. Overall experimental results revealed that about 18% copied text included in General Topics domain and 15% copied text included in Computer Literature domain in CLUE corpus. It was also found that plagiarism ratio was higher in files that were translated from Google translator. The average ratio of plagiarism in both domains when translated from Bing translator was 7.5% and 25.5% when translated from Google translator.

## REFERENCE

Ahmadzia, H.K., M. Patel and S. Coleman (2015). Obstetric surgical site infections: 2 grams compared with 3 grams of cefazolin in morbidly obese women. Obst. Gyn. 126(4) 708-715.

Ali, A.M.E.T. and H.M.D Abdulla (2011). Overview and Comparison of Plagiarism Detection Tools. DATESO 161-172

Alzahrani, S. and N. Salim (2010). Fuzzy semantic-based string similarity for extrinsic plagiarism detection. Bras. H. 1176, 1-8.

Asghari, H., K. Khoshnava and H. Faili (2015). Developing bilingual plagiarism detection

corpus using sentence aligned parallel corpus. Notebook for PAN at CLEF.

Asghari, H. and P. Rosso (2016). Algorithms and corpora for Persian plagiarism detection. F. IRE Spr. Cham. (61-79).

Barrón-Cedeno, A. and P. Rosso (2010). Plagiarism detection across distant language pairs. Asso. Com. Ling. 37-45

Barrón-Cedeño, A., and P. Rosso (2013). Methods for cross-language plagiarism detection. Know. B. Stms. 50, 211-217.

Barrón-Cedeño, A. and P. Rosso (2013). Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Com. Ling. 39(4) 917-947.

Buntinx, V. and F. Kaplan (2015). Inversed N-gram viewer: Searching the space of word temporal profiles. Dig. Hum. 2015 (No. EPFL-CONF-210392).

Cheema, W.A. and S.H. Sittar (2015). A Corpus for Analyzing Text Reuse by People of Different Groups. CLEF.

Chiu, S. and I. Uysal (2010), August. Evaluating text reuse discovery on the web. In Proceedings of the third symposium on Information interaction in context 299-304

Chong, M. and L. Specia (2010). Using natural language processing for automatic detection of plagiarism. IPC.

Clough, P. and M. Stevenson (2011). Developing a corpus of plagiarised short answers. Lang. R. Eval. 45(1) 5-24.

Forstall, C. and S. Jacobson (2014). Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. Dig. S. Hum. 30(4) 503-515.

Danilova, V. (2013). Cross-language plagiarism detection methods. RANLP 2013 51-57

Ekbal, A. and S. Saha (2012), December. Plagiarism detection in text using vector space model. 12th Int. Conf. IEEE. (2012) 366-371

Evangelopoulos, N.E., (2013). Latent semantic analysis. Cog. Sci, 4(6) 683-692.

Ferrero, J. and F. Agnes (2017). Deep Investigation of Cross-Language Plagiarism Detection Methods. arXiv:1705.08828.

Field, H. and R. MR Coulson (2013). An automated graphics tool for comparative genomics: the Coulson plot generator. BMC bio. Info. 14(1) 141.

Flores, E. and P. Rosso, (2015). Cross-Language Source Code Re-Use Detection Using Latent Semantic Analysis. J. UCS. 21(13) 1708-1725.

Forstall, C. and S. Jacobson (2014). Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching. Dig. Sch. Hum. 30(4) 503-515.

Franco-Salvador, P., (2014). Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In Brid. IRD 227-236

Franco-Salvador, M. and P. Rosso (2013), March. Cross-language plagiarism detection using a multilingual semantic network. In Eu. Conf. on IR 710-713.

Gopalakrishnan, S., (2013). Authorship Attribution Based on Grammar Signatures. DDUC.

Gupta, P., and P. Rosso (2011). Detection of Paraphrastic Cases of Mono-lingual and Cross-lingual Plagiarism. ICON.

Gupta, P. and P. Rosso (2012), July. Text reuse with ACL :( upward) trends. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. Assoc. Com. Ling. 76-82.

Hanif, I. and A. Arbab (2015). Cross-language Urdu–English (clue) text alignment corpus. CLEF.

Jänicke, S. and G. Scheuermann (2014), January. Visualizations for text re-use. In Information Visualization Theory and Applications. Int. conf IEEE.59-70).

Kent, C.K. and N. Salim, (2010), September. Web based cross language plagiarism detection. In Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 IEEE. 199-204.

Mahmoodi, M. and M. M Varnamkhasti, (2014). Design a Persian Automated Plagiarism Detector (AMZPPD). arXiv.

Markov, I. and E. Stamatatos, E. and G. Sidorov (2017). Improving cross-topic authorship attribution: The role of pre-processing. CICLing 2017.

Mitkov, R., (2016). Computational Phraseology light: automatic translation of multiword expressions without translation resources. Y.B. Phras.

Nadil, M., F. Souami and H. Sahbi (2016). KCCA-based technique for profile face identification. EURASIP, 2017(1) 2.

Oberreuter, G. and J.D Velásquez (2011). Approaches for intrinsic and external plagiarism detection. Proc. of PAN.

Omar, K. and M. Dashash (2013). The implementation of plagiarism detection system in health sciences publications in Arabic and English languages. I. RE. CO. S., 8, 915-919.

Park, H.M., (2015). Univariate analysis and normality test using SAS, Stata, and SPSS. IU Scholar.

Potthast, M. and P. Rosso (2011). Cross-language plagiarism detection. L.R Eval. 45(1), 45-62.

Potthast, M., P. Rosso and B. Stein (2013). Overview of the 5th international competition on plagiarism detection. CELCT 301-331.

Rabiu, I. and N. Salim (2014). Textual and Structural Approaches to Detecting Figure Plagiarism in Scientific Publications. J. heor, AIT, 70(2).

Rana, M. and U. Babar (2016). A Textual Description Based Approach to Process Matching. In IFIP 194-208.

Saffari, M., S. Sajjadi and M. Mohammadi (2017). Evaluation of Machine Translation (Google Translate vs. Bing Translator) from English into Persian across Academic Fields. Mod. J. LTM. 7(8), 429-442.

Sapkota, U., S. Bethard and T. Solorio (2015). Not all character n-grams are created equal: A study in authorship attribution. Assoc. comp. ling. HLT 93-102.

Seaward, L. and S. Matwin (2009), September. Intrinsic plagiarism detection using complexity analysis. Proc. SEPLN. 56-61.

Shen, Y., L. Deng and G. Mesnil (2014), November. A latent semantic model with convolutional-pooling structure for information retrieval. ACM I.C on IKM 101-110.

Souilem, N., I. Elaissi and H. Messaoud (2017). On the use of KPCA pre-filtering for KCCA method. Inter. J. AMT. 91(9-12), 4331-4340.

Stapleton, P., (2012). Gauging the effectiveness of anti-plagiarism software: An empirical study of second language graduate writers. J. of Eng for Aca, 11(2), 125-133.

Vani, K. and D. Gupta (2017). Text plagiarism classification using syntax based linguistic features. ESA, 88, 448-464.

Vinod, K.R., S. Sandhya and A. Harani (2011). Plagiarism-history, detection and prevention. Hyg. JD, 1-4.

Xiao, M. and Y. Guo (2014), July. Semi-Supervised Matrix Completion for Cross-Lingual Text Classification. AAAI 1607-1614.